

B3

## (12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property  
Organization  
International Bureau



(43) International Publication Date  
2 September 2004 (02.09.2004)

PCT

(10) International Publication Number  
**WO 2004/074436 A2**

- (51) International Patent Classification<sup>7</sup>: C12N (81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.
- (21) International Application Number: PCT/US2004/004060
- (22) International Filing Date: 11 February 2004 (11.02.2004)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data: 60/448,959 19 February 2003 (19.02.2003) US
- (71) Applicant (for all designated States except US): INCYTE CORPORATION [US/US]; 3160 Porter Drive, Palo Alto, CA 94304 (US).
- (72) Inventor; and
- (75) Inventor/Applicant (for US only): LASEK, Amy, W. [US/US]; 297 41st #5, Oakland, CA 94611 (US).
- (74) Agents: HAMLET-COX, Diana et al.; Incyte Corporation, 3160 Porter Drive, Palo Alto, CA 94304 (US).
- (84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).
- Published:  
— without international search report and to be republished upon receipt of that report
- For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: METHODS OF USE OF A GPCR IN THE DIAGNOSIS AND TREATMENT OF COLON AND LUNG CANCER

5'CC	11	20	29	38	47	56	
CCT	ACT	TCG	GGC	ACC	ATG	GAC	ACC
---	---	---	---	---	---	---	---
	Met	Asp	Thr	Ser	Arg	Leu	Gly
	Val	Leu	Leu	Ser	Leu	Leu	Pro
	65	74	83	92	101	110	
GTG	CTG	CTG	CAG	CTG	GCG	ACC	GGG
---	---	---	---	---	---	---	---
Val	Leu	Leu	Gln	Leu	Ala	Thr	Gly
	119	128	137	146	155	164	
AGG	GGC	TGC	CCC	ACA	CAC	TGT	CAT
---	---	---	---	---	---	---	---
Arg	Gly	Cys	Pro	Thr	His	Cys	His
	173	182	191	200	209	218	
GTG	GAC	TGC	TCC	GAC	CTG	GGG	CTC
---	---	---	---	---	---	---	---
Val	Asp	Cys	Ser	Asp	Leu	Gly	Leu
	227	236	245	254	263	272	
ACC	TCC	TAC	CTA	GAC	CTC	AGT	ATG
---	---	---	---	---	---	---	---
Thr	Ser	Tyr	Leu	Asp	Leu	Ser	Met
	281	290	299	308	317	326	
CTG	CCC	AGT	CTC	CGC	TTC	CTG	GAG
---	---	---	---	---	---	---	---
Leu	Pro	Ser	Leu	Arg	Phe	Leu	Glu

(57) Abstract: The invention provides for the use of a protein, HG38, a polynucleotide encoding the protein, and antibodies that specifically bind the protein in various methods to diagnose, stage, treat, or monitor the treatment of colon and lung cancer.

WO 2004/074436 A2

## METHODS OF USE OF A GPCR IN THE DIAGNOSIS AND TREATMENT OF COLON AND LUNG CANCER

5

### TECHNICAL FIELD

This invention relates to the use of a G protein-coupled receptor (GPCR) expressed in colon and lung cancer, its encoding polynucleotide, and an antibody that specifically binds the protein to diagnose, to stage, to treat, or to monitor the progression or treatment of colon or lung cancer.

### BACKGROUND OF THE INVENTION

10 Cancers and malignant tumors are characterized by continuous cell proliferation and cell death and are related causally to both genetics and the environment. Genes whose expression are associated with cancer are of potentially great importance as cancer markers in the early diagnosis and prognosis of various cancers, as well as potential targets in cancer treatment.

Colorectal cancer is the fourth most common cancer and the second most common cause of  
15 cancer death in the United States with approximately 130,000 new cases and 55,000 deaths per year. Colon and rectal cancers share many environmental risk factors and both are found in individuals with specific genetic syndromes. (See Potter (1999) J Natl Cancer Institute 91:916-932 for a review of colorectal cancer.) Colon cancer is the only cancer that occurs with approximately equal frequency in men and women, and the five-year survival rate following diagnosis of colon cancer is  
20 around 55% in the United States (Ries *et al.* (1990) National Institutes of Health, DHHS Publ No. (NIH)90-2789).

Lung cancer is the leading cause of cancer death in the United States affecting more than 100,000 men and 50,000 women each year, and nearly 90% of those diagnosed with lung cancer are cigarette smokers. Tobacco smoke contains substances that induce carcinogen metabolizing enzymes  
25 and covalent DNA adduct formation in exposed bronchial epithelium. In nearly 80% of those diagnosed, the lung cancer has metastasized to pleura, brain, bone, pericardium, and liver. Treatment with surgery, radiation therapy, or chemotherapy is made on the basis of tumor histology, response to growth factors or hormones, and sensitivity to inhibitors or drugs. With current treatments, most patients die within one year of diagnosis. Earlier diagnosis and a systematic approach to the  
30 identification, staging, and treatment of lung cancer could positively affect prognosis.

G protein coupled receptors (GPCRs) are a superfamily of seven transmembrane domain proteins that mediate the transduction of extracellular signals across the plasma membrane of cells from a large, diverse number of ligands through interaction with heterotrimeric G proteins. See, e.g., Watson, S. and S. Arkinstall (1994) The G-protein Linked Receptor Facts Book, Academic Press,  
35 San Diego CA, pp. 2-6.

- The glycoprotein hormone receptors are a subfamily of GPCRs activated by the gonadotropins; lutropin (LH), thyrotropin (TSH), follitropin (FSH) and human choriogonadotropin (hCG), essential for the growth and differentiation of the gonads and thyroid gland. The glycoprotein hormone GPCRs are characterized by a large N-terminal extracellular domain containing some 9 leucine-  
5 repeat domains that function in protein:protein interactions and are likely important for interaction with their respective protein ligands. Two orphan GPCRs (ligand currently unknown) related to this subfamily, designated HG38 and LGR5, have also been identified which share ~35% overall identity with members of the glycoprotein hormone receptor subfamily (McDonald *et al.* (1998) *Biochem Biophys Res Commun* 247:266-270; Hsu *et al.* (1998) *Molecular Endocrinology* 12:1830-1845).  
10 Both receptors are expressed primarily in skeletal muscle, placenta, spinal cord, and various regions of the brain. Like other members of the glycoprotein hormone receptor family, both receptors are characterized by a large extracellular domain containing leucine-rich repeats and a cysteine-rich domain near the junction of the extracellular domain and the first transmembrane domain (Hsu *et al. supra*).  
15 Array technologies and quantitative PCR provide the means to explore the expression profiles of a large number of related or unrelated genes. When an expression profile is examined, arrays provide a platform for examining which genes are tissue-specific, carrying out housekeeping functions, parts of a signaling cascade, or specifically related to a particular genetic predisposition, condition, disease, or disorder. The application of expression profiling is particularly relevant to  
20 improving diagnosis, prognosis, and treatment of the disease.

The discovery of a GPCR and its encoding polynucleotide that are differentially expressed in colon and lung cancer, satisfies a need in the art by providing compositions which are useful to diagnose, to stage, to treat, or to monitor the progression or treatment of colon and lung cancer.

25

#### SUMMARY OF THE INVENTION

The invention is based on the discovery that a GPCR known as HG38 (SEQ ID NO:1) is differentially expressed in colon and lung cancer, and to the use of the protein, its encoding polynucleotide or the complement thereof, and an antibody that specifically binds the protein to  
diagnose, to stage, to treat, or to monitor the progression or treatment of a colon or lung cancer.

30

The invention provides a method for using a polynucleotide to detect the differential expression of a nucleic acid in a sample of colon or lung tissue comprising hybridizing a probe to the nucleic acids, thereby forming hybridization complexes and comparing hybridization complex formation with a standard, wherein the comparison indicates the differential expression of the polynucleotide in the sample. In one aspect, the method of detection further comprises amplifying

the nucleic acids of the sample prior to hybridization. In another aspect, the method showing differential expression of the polynucleotide is used to diagnose a colon or lung cancer.

The invention provides a purified protein or a portion thereof comprising an amino acid sequence of SEQ ID NO:1 for use in the diagnosis of a colon or lung cancer. The invention  
5 provides a method for diagnosing a colon or lung cancer comprising performing an assay to quantify the amount of the protein expressed in a sample of colon or lung tissue and comparing the amount of protein expressed to a standard, thereby diagnosing a colon or lung cancer. In a one aspect, the assay is selected from antibody arrays, enzyme-linked immunosorbent assays, fluorescence-activated cell sorting, 2D-PAGE and scintillation counting, protein arrays, radioimmunoassays, and western  
10 analysis.

The invention also provides a method for using an antibody to detect differential expression of a protein in a sample of colon or lung tissue, the method comprising combining the antibody with a sample under conditions for formation of antibody:protein complexes, detecting complex formation, comparing complex formation with a standard, wherein differential expression of the protein between  
15 the sample and the standard is diagnostic of a colon or lung cancer.

The invention further provides an antagonist which specifically binds the protein having the amino acid sequence of SEQ ID NO:1 for use in the treatment of a colon or lung cancer.

The invention provides a method for treating a colon or lung cancer comprising administering to a subject in need of therapeutic intervention an antibody that specifically binds the protein or a  
20 composition comprising an antibody and a pharmaceutical agent. The invention also provides a method for delivering a pharmaceutical or therapeutic agent to a colon cancer cell comprising attaching the pharmaceutical or therapeutic agent to an antibody that specifically binds the protein and administering the antibody to a subject in need of therapeutic intervention, wherein the antibody delivers the pharmaceutical or therapeutic agent to the cell.

25 The invention also provides a method for using a polynucleotide to produce a mammalian model system, the method comprising constructing a vector containing the polynucleotide of SEQ ID NO:3, transforming the vector into an embryonic stem cell, selecting a transformed embryonic stem cell, microinjecting the transformed embryonic stem cell into a mammalian blastocyst, thereby forming a chimeric blastocyst, transferring the chimeric blastocyst into a pseudopregnant dam,  
30 wherein the dam gives birth to a chimeric offspring containing the polynucleotide in its germ line, and breeding the chimeric mammal to produce a homozygous, mammalian model system.

### BRIEF DESCRIPTION OF THE FIGURES AND TABLES

Figures 1A-1J show the amino acid sequence (SEQ ID NO:1) of HG38 encoded by the nucleic acid sequence of SEQ ID NO:2. The alignment was produced using MACDNASIS PRO software (Hitachi Software Engineering, San Bruno CA).

5 Figure 2 shows the relative expression of HG38 in various normal adult tissues. The X-axis lists tissue type, and the Y-axis, the relative expression of HG38 normalized to that found in normal colon. QPCR analysis was performed using the TAQMAN protocol (Applied Biosystems (ABI), Foster City CA). Tissues were obtained from Clinomics (Pittsfield MA) and Clontech (Palo Alto CA).

10 Figures 3 and 4 show the differential expression of HG38 in donor-matched normal/tumor colon samples as determined using QPCR (ABI). Figure 3 tissues were obtained from the Huntsman Cancer Institute (HCI; Salt Lake City UT); Figure 4 tissues, from Asterand Bioresources, Inc. (Detroit MI).

Figure 5 shows the relative expression of HG38 in colon tumor cell lines compared to a  
15 normal colon cell line (LS123) using QPCR (ABI). Cell lines were obtained from the ATCC (Manassas VA).

Figure 6 shows the differential expression of HG38 in donor-matched normal/tumor lung samples as determined using QPCR (ABI). Tissue samples were obtained from the Roy Castle Institute for Lung Cancer Research (RCI; Liverpool UK).

20 The probe sequence for the QPCR analyses depicted in Figures 2-6 used an oligonucleotide extending from about nucleotide 274 to about nucleotide 291 of SEQ ID NO:2.

Figure 7 shows the expression of the transcript encoding HG38 in normal colon tissue. Thin sections were stained with DAPI and hybridized in situ using sense or antisense RNA probes made from a fragment of SEQ ID NO:2 extending from about nucleotide 274 to about nucleotide 2724 of  
25 SEQ ID NO:2.

Figure 8 shows the expression of the transcript encoding HG38 in a villous adenocarcinoma of the colon. Thin sections were stained with DAPI and hybridized in situ using sense or antisense RNA probes made from a fragment of SEQ ID NO:2 extending from about nucleotide 274 to about nucleotide 2724 of SEQ ID NO:2.

30 Table 1 shows the differential expression of HG38 in cancerous colon tissue relative to normal colon tissue as determined by microarray analysis. Column 1 shows the differential expression of HG38 in terms of the ratio of the signal intensity for the fluorescent dye Cy5 in labeled tumor tissue relative to that for fluorescent dye Cy3 in labeled normal colon tissue. Column 2 shows the source of the normal colon tissue as the individual donor (Dn), or pooled tissue from more than

one donor (pool), column 3 is a description of the colon tumor sample, and column 4, the source of the tumor sample. Tissue samples were obtained from HCI and Asterand Bioresources.

Table 2 shows the differential expression of HG38 in cancerous lung tissue relative to normal lung tissue as determined by microarray analysis. Column headings are the same as those described above for Table 1. Tissue samples were obtained from RCI.

#### DESCRIPTION OF THE INVENTION

It is understood that this invention is not limited to the particular machines, materials and methods described. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments and is not intended to limit the scope of the present invention which will be limited only by the appended claims. As used herein, the singular forms "a", "an", and "the" may include plural reference unless the context clearly dictates otherwise. For example, a reference to "a host cell" includes a plurality of such host cells known to those skilled in the art.

Unless defined otherwise, all technical and scientific terms used herein have the same meanings as commonly understood by one of ordinary skill in the art to which this invention belongs. All publications mentioned herein are cited for the purpose of describing and disclosing the cell lines, protocols, reagents and vectors which are reported in the publications and which might be used in connection with the invention. Nothing herein is to be construed as an admission that the invention is not entitled to antedate such disclosure by virtue of prior invention.

#### Definitions

"Antibody" refers to intact immunoglobulin molecule, a polyclonal antibody, a monoclonal antibody, a chimeric antibody, a recombinant antibody, a humanized antibody, single chain antibodies, a Fab fragment, an F(ab')<sub>2</sub> fragment, an Fv fragment, and an antibody-peptide fusion protein.

"Antigenic determinant" refers to an antigenic or immunogenic epitope, structural feature, or region of an oligopeptide, peptide, or protein which is capable of inducing formation of an antibody that specifically binds the protein. Biological activity is not a prerequisite for immunogenicity.

"Array" refers to an ordered arrangement of at least two polynucleotides, proteins, or antibodies on a substrate. At least one of the polynucleotides, proteins, or antibodies represents a control or standard, and the other polynucleotide, protein, or antibody is of diagnostic or therapeutic interest. The arrangement of at least two and up to about 40,000 polynucleotides, proteins, or antibodies on the substrate assures that the size and signal intensity of each labeled complex, formed between each polynucleotide and at least one nucleic acid, each protein and at least one ligand or antibody, or each antibody and at least one protein to which the antibody specifically binds, is individually distinguishable.

"HG38" refers to a GPCR that is exactly or highly homologous (>85%) to the amino acid sequence of SEQ ID NO:1 obtained from any species including bovine, ovine, porcine, murine, equine,

and preferably the human species, and from any source, whether natural, synthetic, semi-synthetic, or recombinant.

The "complement" of a polynucleotide of the Sequence Listing refers to a nucleic acid molecule which is completely complementary over its full length and which will hybridize to a nucleic acid  
5 molecule under conditions of high stringency.

The phrase "polynucleotide encoding a protein" refers to a nucleic acid whose sequence closely aligns with sequences that encode conserved regions, motifs or domains identified by employing analyses well known in the art. These analyses include BLAST (Basic Local Alignment Search Tool; Altschul (1993) J Mol Evol 36:290-300; Altschul *et al.* (1990) J Mol Biol 215:403-410) and BLAST2 (Altschul *et al.* (1997) Nucleic Acids Res 25:3389-3402) which provide identity within the conserved region. Brenner  
10 *et al.* (1998; Proc Natl Acad Sci 95:6073-6078) who analyzed BLAST for its ability to identify structural homologs by sequence identity found 30% identity is a reliable threshold for sequence alignments of at least 150 residues and 40% is a reasonable threshold for alignments of at least 70 residues (Brenner, page 6076, column 2).

15 A "composition comprising a given polynucleotide" and a "composition comprising a given polypeptide" can refer to any composition containing the given polynucleotide or polypeptide. The composition may comprise a dry formulation or an aqueous solution. Compositions comprising polynucleotides encoding HG38 or fragments of HG38 may be employed as hybridization probes. The probes may be stored in freeze-dried form and may be associated with a stabilizing agent such as a  
20 carbohydrate. In hybridizations, the probe may be deployed in an aqueous solution containing salts (e.g., NaCl), detergents (e.g., sodium dodecyl sulfate; SDS), and other components (e.g., Denhardt's solution, dry milk, salmon sperm DNA, etc.).

A "deletion" refers to a change in the amino acid or nucleotide sequence that results in the absence of one or more amino acid residues or nucleotides.

25 "Derivative" refers to a polynucleotide or a protein that has been subjected to a chemical modification. Derivatization of a polynucleotide can involve substitution of a nontraditional base such as queosine or of an analog such as hypoxanthine. These substitutions are well known in the art. Derivatization of a polynucleotide or a protein can also involve the replacement of a hydrogen by an acetyl, acyl, alkyl, amino, formyl, or morpholino group (for example, 5-methylcytosine). Derivative  
30 molecules retain the biological activities of the naturally occurring molecules but may confer longer lifespan or enhanced activity.

"Differential expression" refers to increased or upregulated; or decreased, downregulated, or absent gene or protein expression, determined by comparing at least two different samples. Such

comparisons may be carried out between, for example, a treated and an untreated sample, or a diseased and a normal sample.

"Disorder" refers to conditions, diseases or syndromes in which HG38 or the mRNA encoding HG38 are differentially expressed; in particular, a colon or lung cancer.

5       An "expression profile" is a representation of gene expression in a sample. A nucleic acid expression profile is produced using sequencing, hybridization, or amplification (quantitative PCR) technologies and mRNAs or polynucleotides from a sample. A protein expression profile, although time delayed, mirrors the nucleic acid expression profile and may use antibody or protein arrays, enzyme-linked immunosorbent assays, fluorescence-activated cell sorting, spatial immobilization such as 2D-PAGE, and  
10 radioimmunoassays including radiolabeling and quantification using a scintillation counter and western analysis to detect protein expression in a sample. The nucleic acids, proteins, or antibodies may be used in solution or attached to a substrate, and their detection is based on methods and labeling moieties well known in the art. Expression profiles may also be evaluated by methods such as electronic northern analysis, guilt-by-association, and transcript imaging. Expression profiles produced using any of the  
15 above methods may be contrasted with expression profiles produced using normal or diseased tissues. Of note is the correspondence between mRNA and protein expression has been discussed by Zweiger (2001, Transducing the Genome, McGraw-Hill, San Francisco, CA) and Glavas et al. (2001; T cell activation upregulates cyclic nucleotide phosphodiesterases 8A1 and 7A3, Proc Natl Acad Sci 98:6319-6342) among others.

20       The term "hybridization complex" refers to a complex formed between two nucleic acids by virtue of the formation of hydrogen bonds between complementary bases. A hybridization complex may be formed in solution (e.g., C<sub>0</sub>t or R<sub>0</sub>t analysis) or formed between one nucleic acid present in solution and another nucleic acid immobilized on a solid support (e.g., paper, membranes, filters, chips, pins or glass slides, or any other appropriate substrate to which cells or their nucleic acids have been fixed).

25       "Identity" as applied to sequences, refers to the quantification (usually percentage) of nucleotide or residue matches between at least two sequences aligned using a standardized algorithm such as Smith-Waterman alignment (Smith and Waterman (1981) J Mol Biol 147:195-197), CLUSTALW (Thompson et al. (1994) Nucleic Acids Res 22:4673-4680), or BLAST2 (Altschul (1997, supra). BLAST2 may be used in a standardized and reproducible way to insert gaps in one of the sequences in order to optimize  
30 alignment and to achieve a more meaningful comparison between them. "Similarity" uses the same algorithms but takes conservative substitution of residues into account. In proteins, similarity exceeds identity in that substitution, for example, of a valine for a leucine or isoleucine, is counted in calculating the reported percentage. Substitutions which are considered to be conservative are well known in the art.



An "immunogenic fragment" is a polypeptide or oligopeptide fragment of HG38 which is capable of eliciting an immune response when introduced into a living organism, for example, a mammal. The term "immunogenic fragment" also includes any polypeptide or oligopeptide fragment of HG38 which is useful in any of the antibody production methods disclosed herein or known in the art.

5 "Labeling moiety" refers to any reporter molecule including radionuclides, enzymes, fluorescent, chemiluminescent, or chromogenic agents, substrates, cofactors, inhibitors, or magnetic particles than can be attached to or incorporated into a polynucleotide, protein, or antibody. A wide variety conjugation techniques are known in the art and include both direct synthesis and chemical conjugation, particularly to amines, thiols and other side groups which may be present. Visible labels and dyes include but are not  
10 limited to anthocyanins,  $\beta$  glucuronidase, biotin, BIODIPY, Coomassie blue, Cy3 and Cy5, 4,6-diamidino-2-phenylindole (DAPI), digoxigenin, fluorescein, FITC, gold, green fluorescent protein (GFP), lissamine, luciferase, phycoerythrin, rhodamine, spiro red, silver, streptavidin, and the like. Radioactive markers include radioactive forms of hydrogen, iodine, phosphorous, sulfur, and the like.

"Ligand" refers to any agent, molecule, or compound which will bind specifically to a  
15 polynucleotide or to an epitope of a protein. Such ligands stabilize or modulate the activity of polynucleotides or proteins and may be composed of inorganic and/or organic substances including minerals, cofactors, nucleic acids, proteins, carbohydrates, fats, and lipids.

The term "microarray" refers to an arrangement of a plurality of polynucleotides, polypeptides, antibodies, or other chemical compounds on a substrate.

20 The terms "element" and "array element" refer to a polynucleotide, polypeptide, antibody, or other chemical compound having a unique and defined position on a microarray.

The term "modulate" refers to a change in the activity of HG38. For example, modulation may cause an increase or a decrease in protein activity, binding characteristics, or any other biological, functional, or immunological properties of HG38.

25 A "multispecific molecule" can bind with at least two different binding specificities to at least two different molecules or two different sites on a molecule. Antibodies can perform as multispecific molecules in that they can bind to both a target protein and a pharmaceutical agent.

The phrases "nucleic acid" and "nucleic acid sequence" refer to a nucleotide, oligonucleotide, polynucleotide, or any fragment thereof. These phrases also refer to DNA or RNA of genomic or  
30 synthetic origin which may be single-stranded or double-stranded and may represent the sense or the antisense strand, to peptide nucleic acid (PNA), or to any DNA-like or RNA-like material.

"Oligonucleotide" refers a single-stranded molecule from about 18 to about 60 nucleotides in length which may be used in hybridization or amplification technologies or in regulation of replication, transcription or translation. Equivalent terms are amplicon, ampimer, primer, and oligomer.

"Operably linked" refers to the situation in which a first nucleic acid sequence is placed in a functional relationship with a second nucleic acid sequence. For instance, a promoter is operably linked to a coding sequence if the promoter affects the transcription or expression of the coding sequence.

Operably linked DNA sequences may be in close proximity or contiguous and, where necessary to join  
5 two protein coding regions, in the same reading frame.

"Peptide nucleic acid" (PNA) refers to an antisense molecule or anti-gene agent which comprises an oligonucleotide of at least about 5 nucleotides in length linked to a peptide backbone of amino acid residues ending in lysine. The terminal lysine confers solubility to the composition. PNAs preferentially bind complementary single stranded DNA or RNA and stop transcript elongation, and may be pegylated to  
10 extend their lifespan in the cell.

A "pharmaceutical agent" may be an antibody, an antisense molecule, a bispecific molecule, a multispecific molecule, a peptide, a protein, a radionuclide, a small drug molecule, a cytospecific or cytotoxic drug such as abrin, actinomyosin D, cisplatin, croton, doxorubicin, 5-fluorouracil, methotrexate, ricin, vincristine, vinblastine, or any combination of these elements.

15 "Post-translational modification" of a protein can involve lipidation, glycosylation, phosphorylation, acetylation, racemization, proteolytic cleavage, and the like. These processes may occur synthetically or biochemically. Biochemical modifications will vary by cellular location, cell type, pH, enzymatic milieu, and the like.

"Probe" refers to polynucleotides encoding HG38, their complements, or fragments thereof, which  
20 are used to detect identical, allelic or related polynucleotides. Probes are isolated oligonucleotides or polynucleotides attached to a detectable label or reporter molecule. Typical labels include radioactive isotopes, ligands, chemiluminescent agents, and enzymes. "Primers" are short nucleic acids, usually DNA oligonucleotides, which may be annealed to a target polynucleotide by complementary base-pairing. The primer may then be extended along the target DNA strand by a DNA polymerase enzyme. Primer pairs  
25 can be used for amplification (and identification) of a nucleic acid, e.g., by the polymerase chain reaction (PCR).

"Protein" refers to a polypeptide or any portion thereof. A "portion" of a protein refers to that length of amino acid sequence which would retain at least one biological activity, a domain identified by PFAM or PRINTS analysis or an antigenic determinant of the protein identified using Kyte-Doolittle  
30 algorithms of the PROTEAN program (DNASTAR, Madison WI). An "oligopeptide" is an amino acid sequence from about five residues to about 15 residues that is used as part of a fusion protein to produce an antibody.

A "recombinant nucleic acid" is a nucleic acid that is not naturally occurring or has a sequence that is made by an artificial combination of two or more otherwise separated segments of sequence. This

artificial combination is often accomplished by chemical synthesis or, more commonly, by the artificial manipulation of isolated segments of nucleic acids, e.g., by genetic engineering techniques such as those described in Sambrook and Russell (*supra*). The term recombinant includes nucleic acids that have been altered solely by addition, substitution, or deletion of a portion of the nucleic acid. Frequently, a  
5 recombinant nucleic acid may include a nucleic acid sequence operably linked to a promoter sequence. Such a recombinant nucleic acid may be part of a vector that is used, for example, to transform a cell.

Alternatively, such recombinant nucleic acids may be part of a viral vector, e.g., based on a vaccinia virus, that could be used to vaccinate a mammal wherein the recombinant nucleic acid is expressed, inducing a protective immunological response in the mammal.

10 A "regulatory element" refers to a nucleic acid sequence usually derived from untranslated regions of a gene and includes enhancers, promoters, introns, and 5' and 3' untranslated regions (UTRs). Regulatory elements interact with host or viral proteins which control transcription, translation, or RNA stability.

"Reporter molecules" are chemical or biochemical moieties used for labeling a nucleic acid,  
15 amino acid, or antibody. Reporter molecules include radionuclides; enzymes; fluorescent, chemiluminescent, or chromogenic agents; substrates; cofactors; inhibitors; magnetic particles; and other moieties known in the art.

An "RNA equivalent," in reference to a DNA molecule, is composed of the same linear sequence of nucleotides as the reference DNA molecule with the exception that all occurrences of the nitrogenous  
20 base thymine are replaced with uracil, and the sugar backbone is composed of ribose instead of deoxyribose.

"Sample" is used in its broadest sense as containing nucleic acids, proteins, and antibodies. A sample may comprise a bodily fluid such as ascites, blood, cerebrospinal fluid, lymph, semen, sputum, urine and the like; the soluble fraction of a cell preparation, or an aliquot of media in which cells were  
25 grown; a chromosome, an organelle, or membrane isolated or extracted from a cell; genomic DNA, RNA, or cDNA in solution or bound to a substrate; a cell; a tissue, a tissue biopsy, or a tissue print; buccal cells, skin, hair, a hair follicle; and the like.

The terms "specific binding" and "specifically binding" refer to that interaction between a protein or peptide and an agonist, an antibody, an antagonist, a small molecule, or any natural or synthetic binding  
30 composition. The interaction is dependent upon the presence of a particular structure of the protein, e.g., the antigenic determinant or epitope, recognized by the binding molecule. For example, if an antibody is specific for epitope "A," the presence of a polypeptide comprising the epitope A, or the presence of free unlabeled A, in a reaction containing free labeled A and the antibody will reduce the amount of labeled A that binds to the antibody.

The term "substantially purified" refers to nucleic acid or amino acid sequences that are removed from their natural environment and are isolated or separated, and are at least about 60% free, preferably at least about 75% free, and most preferably at least about 90% free from other components with which they are naturally associated.

5 A "substitution" refers to the replacement of one or more amino acid residues or nucleotides by different amino acid residues or nucleotides, respectively.

"Substrate" refers to any rigid or semi-rigid support to which polynucleotides, proteins, or antibodies are bound and includes magnetic or nonmagnetic beads, capillaries or other tubing, chips, fibers, filters, gels, membranes, plates, polymers, slides, wafers, and microparticles with a variety of  
10 surface forms including channels, columns, pins, pores, trenches, and wells.

A "transcript image" (TI) is a profile of gene transcription activity in a particular tissue at a particular time. TI provides assessment of the relative abundance of expressed polynucleotides in the cDNA libraries of an EST database as described in USPN 5,840,484, incorporated herein by reference.

"Transformation" describes a process by which exogenous DNA is introduced into a recipient cell.  
15 Transformation may occur under natural or artificial conditions according to various methods well known in the art, and may rely on any known method for the insertion of foreign nucleic acid sequences into a prokaryotic or eukaryotic host cell. The method for transformation is selected based on the type of host cell being transformed and may include, but is not limited to, bacteriophage or viral infection, electroporation, heat shock, lipofection, and particle bombardment. The term "transformed cells" includes  
20 stable transformed cells in which the inserted DNA is capable of replication either as an autonomously replicating plasmid or as part of the host chromosome, as well as transiently transformed cells which express the inserted DNA or RNA for limited periods of time.

A "transgenic organism," as used herein, is any organism, including but not limited to animals and plants, in which one or more of the cells of the organism contains heterologous nucleic acid introduced by  
25 way of human intervention, such as by transgenic techniques well known in the art. The nucleic acid is introduced into the cell, directly or indirectly by introduction into a precursor of the cell, by way of deliberate genetic manipulation, such as by micro injection or by infection with a recombinant virus. In another embodiment, the nucleic acid can be introduced by infection with a recombinant viral vector, such as a lentiviral vector (Lois, C. et al. (2002) Science 295:868-872). The term genetic manipulation does not  
30 include classical cross-breeding, or *in vitro* fertilization, but rather is directed to the introduction of a recombinant DNA molecule. The transgenic organisms contemplated in accordance with the present invention include bacteria, cyanobacteria, fungi, plants and animals. The isolated DNA of the present invention can be introduced into the host by methods known in the art, for example infection, transfection,

transformation or transconjugation. Techniques for transferring the DNA of the present invention into such organisms are widely known and provided in references such as Sambrook and Russell (*supra*).

"Variant" refers to molecules that are recognized variations of a protein or the polynucleotides that encode it. Splice variants may be determined by BLAST score, wherein the score is at least 100, and most preferably at least 400. Allelic variants have a high percent identity to the polynucleotides and may differ by about three bases per hundred bases. "Single nucleotide polymorphism" (SNP) refers to a change in a single base as a result of a substitution, insertion or deletion. The change may be conservative (purine for purine) or non-conservative (purine to pyrimidine) and may or may not result in a change in an encoded amino acid or its secondary, tertiary, or quaternary structure.

## 10 THE INVENTION

The invention is based on a GPCR (HG38) and its encoding polynucleotide that are differentially expressed in colon and lung cancer, and to the use of the polynucleotide, the protein, and to an antibody that specifically binds the protein in the characterization, diagnosis, prognosis, treatment and evaluation of treatment of colon and lung cancer.

15 Figures 1A-1J shows the amino acid sequence of SEQ ID NO:1 (HG38) encoded by the nucleic acid sequence of SEQ ID NO:2.

Microarray data first showed that SEQ ID NO:2 was preferentially and differentially expressed in colon and lung cancer (Tables 1 and 2, respectively). In particular, the polynucleotide encoding HG38 was overexpressed in 9 of 12 colon tumors (Dn3757, Dn3756, Dn3583, Dn3647, Dn3579, Dn3582, 20 Dn3839, Dn9573, and Dn9576), and 6 of 17 lung tumors (Dn7178, Dn7173, Dn7186, Dn7963, Dn5797, and Dn5796) compared to donor-matched normal colon or lung tissue or, in the case of Dn3757, a pool of normal colon tissue. In these experiments, a value of at least 1.7-fold was considered to be significant differential expression. An average value was considered where duplicate experiments were performed. This discovery led to further studies using QPCR and *in situ* hybridization studies comparing normal and 25 cancerous tissues and tumor cell lines.

Figure 2 shows that HG38 is expressed at the highest levels in brain and skeletal muscle in normal adult tissues, an expression pattern consistent with the known literature.

QPCR analysis of HG38 expression in donor-matched normal/tumor colon samples obtained from the Hunstman Cancer Institute shows that the gene was overexpressed in 8/10 tumor samples; Dn3579, 30 3580, 3581, 3582, 3583, 3647, 3649, and 3479 (Figure 3). A similar study using samples from an alternate source (Asterand Bioresources, Inc) showed overexpression of HG38 in 4/5 colon tumors compared to donor-matched normal colon; Dn9573, Dn9574, Dn9576, and Dn9577 (Figure 4). The highest expression of HG38 was found in one unmatched colon tumor sample, 8401, relative to a pool of normal colon tissue. It is further noteworthy that 6 of the 8 patient samples that exhibited overexpression of HG38 in colon

tumor versus normal colon tissue in the microarray study shown above in Table 1 (Dn3583, 3647, 3579, 3582, 9573 and 9576) likewise showed overexpression of HG38 by QPCR analysis of the same samples. In addition, 2 patient samples that showed marginal overexpression of HG38 in colon tumors in Table 1, Dn9574 and Dn9477, exhibited more significant overexpression by QPCR analysis. Differences in results between the microarray study shown in Table 1 and the QPCR analysis in Figure 3 are likely due, in part, to the greater sensitivity and larger dynamic range for QPCR analysis than for microarray analysis.

Figure 5 shows the expression of the transcript encoding HG38 in colon tumor cell lines relative to a non-tumorigenic colon cell line, LS123, using QPCR. The cell lines were obtained from the ATCC. The highest expression of HG38 was observed in SW620, a metastasis of colon carcinoma derived from ascitic fluid.

Figure 6 shows the expression of the transcript encoding HG38 in donor-matched normal/tumor lung samples obtained from RCI. The gene was overexpressed in 5/15 tumor samples; Dn7173, Dn7178, Dn7191, Dn9751, and Dn9764.

Figure 7 shows expression of the transcript encoding HG38 in the epithelial cells of the colon crypt in normal colon tissue. Transcript expression was visualized using *in situ* hybridization in a thin-sectioned colon sample using sense or antisense RNA probes made from a fragment extending from about nucleotide 274 to about nucleotide 2724 of SEQ ID NO:2. For contrast, the respective sections were counterstained with DAPI.

Figure 8 shows the results of a similar *in situ* hybridization study in a thin-sectioned villous adenocarcinoma colon sample clearly showing expression of the transcript in the tumor epithelium.

Northern analysis conducted using the LIFESEQ GOLD database (Incyte Genomics, Palo Alto, CA) also shows the differential expression of the transcript encoding HG38 in colon and lung tumors. The two tables shown below describe cDNA libraries from colon and lung tissues in which the gene was expressed. The first column shows the library name; the second column, the total number of cDNAs sequenced in that library; the third column, a description of the library; the fourth column, the absolute abundance of the transcript encoding HG38 in the library; the fourth column, the percent abundance of the transcript in the library.

**Category: Digestive System (Colon)**

	<u>Library Name*</u>	<u>cDNAs</u>	<u>Description of Colon Tissue</u>	<u>Abundance</u>	<u>% Abundance</u>
30	COLCDIT03	3069	colon, cecum polyp, aw/adenocA <sup>1</sup> , 67F	3	0.0978
	COLNNOT22	3599	colon mw/Crohn's, 56F	1	0.0278
	COLNTUP17	7417	colon tumor, adenoCA, 3' CGAP	1	0.0135

<sup>1</sup> adenoCA=adenocarcinoma; \* Normalized and fetal libraries were excluded from this analysis.

The data shows that the expression of this transcript in colon tissue libraries is associated exclusively with diseased colon, in particular with colon cancer and colon polyps, a precancerous

condition, and with the inflammatory condition, Crohn's disease. Expression was not found in at least 6 normal adult colon tissues unassociated with disease (COLENOR03, COLENOT01, COLENOT02, COLNNOP01, COLNNOP02, and COLNNOP06).

**Category: Respiratory System (Lung)**

<u>Library Name*</u>	<u>cDNAs</u>	<u>Description of Colon Tissue</u>	<u>Abundance</u>	<u>% Abundance</u>
LUNGTUT17	3950	lung tumor, adenoCA, 53M, m/LUNGNOT28	2	0.0506
LUNGTUT07	3873	lung tumor, squamous cell CA, 50M	1	0.0258

10 \*No tissues were excluded from this analysis

The above data shows the expression of this transcript exclusively in lung cancer tissue.

LUNGTUT17 is particularly significant because it is matched with (m/) normal lung tissue from the same donor (LUNGNOT28) in which expression was undetectable. Expression was also not found in at least 10 normal lung tissue libraries unassociated with disease (LUNGNOE02, LUNGNOM01, LUNGNOP01,

15 LUNGNOT01, LUNGNOT02, LUNGNOT04, LUNGNOT27, LUNGNOT34, LUNGNOT37, and LUNGNOT40).

The differential expression of HG38 in colon and lung cancer tissue relative to normal tissue and, in particular, its localization in epithelial cells of colon tissue provides a basis for the use of the protein, polynucleotides encoding the protein, and antibodies that specifically bind the protein in the detection of  
20 colon and lung cancer, and for the use if the antibody in the treatment of colon or lung cancer either by the delivery of pharmaceutical agents for cancer bound to the antibody, or by the use of the antibody itself as an antagonist of HG38. The use of the receptor itself as a target for antagonists of HG38 in the treatment of colon and lung cancer is also contemplated.

Mammalian variants of the polynucleotide encoding HG38 were identified using BLAST2 with  
25 default parameters and the ZOOSEQ databases (Incyte Genomics). A highly homologous polynucleotide having about 85% identity to the majority of the coding region of the human polynucleotide is shown in the table below. The first column represents the SEQ ID NO: for homologous polynucleotides (SEQ ID<sub>var</sub>); the second column, the Incyte ID for the homologous polynucleotide (Incyte ID<sub>var</sub>); the third column, the species; the fourth column, the percent identity to the human polynucleotide; and the fifth  
30 column, the nucleotide alignment of the homologous polynucleotide to the human polynucleotide.

SEQ ID <sub>var</sub>	Incyte ID <sub>var</sub>	Species	Identity	Nt <sub>H</sub> Alignment
3	050237_Mm.1	Mouse	85%	112-2740

The mammalian polynucleotide of SEQ ID NO:3 may be used in hybridization, amplification, and screening technologies to identify and distinguish among SEQ ID NO:3 and related molecules in a sample.

35 The mammalian polynucleotide, SEQ ID NO:3 may also be used to produce transgenic cell lines or organisms which are model systems for human colon and lung cancer and upon which the toxicity and

efficacy of therapeutic treatments may be tested. Toxicology studies, clinical trials, and subject/patient treatment profiles may be performed and monitored using the polynucleotides, proteins, antibodies and molecules and compounds identified using the polynucleotides and proteins of the present invention.

#### Characterization and Use of the Invention

##### 5 cDNA libraries

mRNA is isolated from mammalian cells and tissues using methods which are well known to those skilled in the art and used to prepare the cDNA libraries. The Incyte cDNAs were isolated from mammalian cDNA libraries prepared as described in the EXAMPLES. The consensus sequence is present in a single clone insert, or chemically assembled, based on the electronic assembly from sequenced  
10 fragments including Incyte polynucleotides and extension and/or shotgun sequences. Computer programs, such as PHRAP (P Green, University of Washington, Seattle WA) and the AUTOASSEMBLER application (ABI), are used in sequence assembly and are described in EXAMPLE V. After verification of the 5' and 3' sequence, at least one representative polynucleotide which encodes HG38 is designated a reagent for research and development.

##### 15 Sequencing

Methods for sequencing nucleic acids are well known in the art and may be used to practice any of the embodiments of the invention. These methods employ enzymes such as the Klenow fragment of DNA polymerase I, SEQUENASE, Taq DNA polymerase and thermostable T7 DNA polymerase (Amersham Biosciences (APB), Piscataway NJ), or combinations of polymerases and proofreading exonucleases  
20 (Invitrogen, Carlsbad CA). Sequence preparation is automated with machines such as the MICROLAB 2200 system (Hamilton, Reno NV) and the DNA ENGINE thermal cycler (MJ Research, Watertown MA) and sequencing, with the PRISM 3700, 377 or 373 DNA sequencing systems (ABI) or the MEGABACE 1000 DNA sequencing system (APB).

The nucleic acid sequences of the polynucleotides presented in the Sequence Listing were  
25 prepared by such automated methods and may contain occasional sequencing errors and unidentified nucleotides, designated with an N, that reflect state-of-the-art technology at the time the polynucleotide was sequenced. Vector, linker, and polyA sequences were masked using algorithms and programs based on BLAST, dynamic programming, and dinucleotide nearest neighbor analysis. Ns and SNPs can be verified either by resequencing the polynucleotide or using algorithms to compare multiple sequences that  
30 overlap the area in which the Ns or SNP occur. Both of these techniques are well known to and used by those skilled in the art. The sequences may be analyzed using a variety of algorithms described in Ausubel et al. (1997; Short Protocols in Molecular Biology, John Wiley & Sons, New York NY, unit 7.7) and in Meyers (1995; Molecular Biology and Biotechnology, Wiley VCH, New York NY, pp. 856-853).



Shotgun sequencing may also be used to complete the sequence of a particular cloned insert of interest. Shotgun strategy involves randomly breaking the original insert into segments of various sizes and cloning these fragments into vectors. The fragments are sequenced and reassembled using overlapping ends until the entire sequence of the original insert is known. Shotgun sequencing methods are well known in the art and use thermostable DNA polymerases, heat-labile DNA polymerases, and primers chosen from representative regions flanking the polynucleotides of interest. Incomplete assembled sequences are inspected for identity using various algorithms or programs such as CONSED (Gordon (1998) Genome Res 8:195-202) which are well known in the art. Contaminating sequences, including vector or chimeric sequences, can be removed, and deleted sequences can be restored to complete the assembled, finished sequences.

#### Extension of a Nucleic Acid Sequence

The sequences of the invention may be extended using various PCR-based methods known in the art. For example, the XL-PCR kit (ABI), nested primers, and cDNA or genomic DNA libraries may be used to extend the nucleic acid sequence. For all PCR-based methods, primers may be designed using software, such as OLIGO primer analysis software (Molecular Biology Insights, Cascade CO) to be about 22 to 30 nucleotides in length, to have a GC content of about 50% or more, and to anneal to a target molecule at temperatures from about 55C to about 68C. When extending a sequence to recover regulatory elements, genomic, rather than cDNA libraries are used.

#### Hybridization

The polynucleotide and fragments thereof can be used in hybridization technologies for various purposes. A probe may be designed or derived from unique regions such as the 5' regulatory region or from a nonconserved region (i.e., 5' or 3' of the nucleotides encoding the conserved catalytic domain of the protein) and used in protocols to identify naturally occurring molecules encoding the HG38, allelic variants, or related molecules. The probe may be DNA or RNA, may be single-stranded, and should have at least 50% sequence identity to any of the nucleic acid sequences, SEQ ID NOs:2 or 3. Hybridization probes may be produced using oligolabeling, nick-translation, end-labeling, or PCR amplification in the presence of a reporter molecule. A vector containing the polynucleotide or a fragment thereof may be used to produce an mRNA probe in vitro by addition of an RNA polymerase and labeled nucleotides. These procedures may be conducted using kits such as those provided by APB.

The stringency of hybridization is determined by G+C content of the probe, salt concentration, and temperature. In particular, stringency can be increased by reducing the concentration of salt or raising the hybridization temperature. Hybridization can be performed at low stringency with buffers, such as 5xSSC with 1% sodium dodecyl sulfate (SDS) at 60C, which permits the formation of a hybridization complex between nucleic acid sequences that contain some mismatches. Subsequent washes are performed at

higher stringency with buffers such as 0.2xSSC with 0.1% SDS at either 45C (medium stringency) or 68C (high stringency). At high stringency, hybridization complexes will remain stable only where the nucleic acids are completely complementary. In some membrane-based hybridizations, from about 35% to about 50% formamide can be added to the hybridization solution to reduce the temperature at which

5 hybridization is performed. Background signals can be reduced by the use of detergents such as Sarkosyl or TRITON X-100 (Sigma-Aldrich) and a blocking agent such as denatured salmon sperm DNA. Selection of components and conditions for hybridization are well known to those skilled in the art and are reviewed in Ausubel (*supra*) and Sambrook *et al.* (1989) Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Press, Plainview NY.

10 Arrays may be prepared and analyzed using methods well known in the art. Oligonucleotides or polynucleotides may be used as hybridization probes or targets to monitor the expression level of large numbers of genes simultaneously or to identify genetic variants, mutations, and single nucleotide polymorphisms. Arrays may be used to determine gene function; to understand the genetic basis of a condition, disease, or disorder; to diagnose a condition, disease, or disorder; and to develop and monitor  
15 the activities of therapeutic agents. (See, e.g., USPN 5,474,796; Schena *et al.* (1996) *Proc Natl Acad Sci* 93:10614-10619; Heller *et al.* (1997) *Proc Natl Acad Sci* 94:2150-2155; USPN 5,605,662.)

Hybridization probes are also useful in mapping the naturally occurring genomic sequence. The probes may be hybridized to a particular chromosome, a specific region of a chromosome, or an artificial chromosome construction. Such constructions include human artificial chromosomes, yeast artificial  
20 chromosomes, bacterial artificial chromosomes, bacterial P1 constructions, or the cDNAs of libraries made from single chromosomes.

#### QPCR

QPCR is a method for quantifying a nucleic acid molecule based on detection of a fluorescent signal produced during PCR amplification (Gibson *et al.* (1996) *Genome Res* 6:995-1001; Heid *et al.*  
25 (1996) *Genome Res* 6:986-994). Amplification is carried out on machines such as the PRISM 7700 detection system (ABI) which consists of a 96-well thermal cycler connected to a laser and charge-coupled device (CCD) optics system. To perform QPCR, a PCR reaction is carried out in the presence of a doubly labeled probe. The probe, which is designed to anneal between the standard forward and reverse PCR primers, is labeled at the 5' end by a fluorescent reporter dye such as 6-carboxyfluorescein (6-FAM) and at  
30 the 3' end by a quencher molecule such as 6-carboxy-tetramethyl-rhodamine (TAMRA). As long as the probe is intact, the 3' quencher extinguishes fluorescence by the 5' reporter. However, during each primer extension cycle, the annealed probe is degraded as a result of the intrinsic 5' to 3' nuclease activity of Taq polymerase (Holland *et al.* (1991) *Proc Natl Acad Sci* 88:7276-7280). This degradation separates the reporter from the quencher, and fluorescence is detected every few seconds by the CCD. The higher the

starting copy number of the nucleic acid, the sooner an increase in fluorescence is observed. A cycle threshold ( $C_T$ ) value, representing the cycle number at which the PCR product crosses a fixed threshold of detection is determined by the instrument software. The  $C_T$  is inversely proportional to the copy number of the template and can therefore be used to calculate either the relative or absolute initial concentration of the nucleic acid molecule in the sample. The relative concentration of two different molecules can be calculated by determining their respective  $C_T$  values (comparative  $C_T$  method). Alternatively, the absolute concentration of the nucleic acid molecule can be calculated by constructing a standard curve using a housekeeping molecule of known concentration. The process of calculating  $C_T$  values, preparing a standard curve, and determining starting copy number is performed using SEQUENCE DETECTOR 1.7 software (ABI).

#### Expression

Any one of a multitude of polynucleotides encoding HG38 may be cloned into a vector and used to express the protein, or portions thereof, in host cells. The nucleic acid sequence can be engineered by such methods as DNA shuffling (USPN 5,830,721) and site-directed mutagenesis to create new restriction sites, alter glycosylation patterns, change codon preference to increase expression in a particular host, produce splice variants, extend half-life, and the like. The expression vector may contain transcriptional and translational control elements (promoters, enhancers, specific initiation signals, and polyadenylated 3' sequence) from various sources which have been selected for their efficiency in a particular host. The vector, polynucleotide, and regulatory elements are combined using in vitro recombinant DNA techniques, synthetic techniques, and/or in vivo genetic recombination techniques well known in the art and described in Sambrook (supra, ch. 4, 8, 16 and 17).

A variety of host systems may be transformed with an expression vector. These include, but are not limited to, bacteria transformed with recombinant bacteriophage, plasmid, or cosmid DNA expression vectors; yeast transformed with yeast expression vectors; insect cell systems transformed with baculovirus expression vectors or plant cell systems transformed with expression vectors containing viral and/or bacterial elements (Ausubel supra, unit 16). In mammalian cell systems, an adenovirus transcriptional/translational complex may be utilized. After sequences are ligated into the E1 or E3 region of the viral genome, the infective virus is used to transform and express the protein in host cells. The Rous sarcoma virus enhancer or SV40 or EBV-based vectors may also be used for high-level protein expression.

Routine cloning, subcloning, and propagation of nucleic acid sequences can be achieved using the multifunctional pBLUESCRIPT vector (Stratagene, La Jolla CA) or pSPORT1 plasmid (Invitrogen). Introduction of a nucleic acid sequence into the multiple cloning site of these vectors disrupts the lacZ gene and allows colorimetric screening for transformed bacteria. In addition, these vectors may be useful

for in vitro transcription, dideoxy sequencing, single strand rescue with helper phage, and creation of nested deletions in the cloned sequence.

For long term production of recombinant proteins, the vector can be stably transformed into cell lines along with a selectable or visible marker gene on the same or on a separate vector. After  
5 transformation, cells are allowed to grow for about 1 to 2 days in enriched media and then are transferred to selective media. Selectable markers, antimetabolite, antibiotic, or herbicide resistance genes, confer resistance to the relevant selective agent and allow growth and recovery of cells which successfully express the introduced sequences. Resistant clones identified either by survival on selective media or by the expression of visible markers may be propagated using culture techniques. Visible markers are also  
10 used to estimate the amount of protein expressed by the introduced genes. Verification that the host cell contains the desired polynucleotide is based on DNA-DNA or DNA-RNA hybridizations or PCR amplification.

The host cell may be chosen for its ability to modify a recombinant protein in a desired fashion. Such modifications include acetylation, carboxylation, glycosylation, phosphorylation, lipidation,  
15 acylation and the like. Post-translational processing which cleaves a "prepro" form may also be used to specify protein targeting, folding, and/or activity. Different host cells which have specific cellular machinery and characteristic mechanisms for post-translational activities may be chosen to ensure the correct modification and processing of the recombinant protein.

#### Recovery of Proteins from Cell Culture

20 Heterologous moieties engineered into a vector for ease of purification include glutathione S-transferase (GST), 6xHis, FLAG, MYC, and the like. GST and 6-His are purified using affinity matrices such as immobilized glutathione and metal-chelate resins, respectively. FLAG and MYC are purified using monoclonal and polyclonal antibodies. For ease of separation following purification, a sequence encoding a proteolytic cleavage site may be part of the vector located between the protein and the  
25 heterologous moiety. Methods for recombinant protein expression and purification are discussed in Ausubel (*supra*, unit 16).

#### Protein Identification

Several techniques have been developed which permit rapid identification of proteins using high performance liquid chromatography and mass spectrometry (MS). Beginning with a sample containing  
30 proteins, the method is: 1) proteins are separated using two-dimensional gel electrophoresis (2-DE), 2) selected proteins are excised from the gel and digested with a protease to produce a set of peptides; and 3) the peptides are subjected to mass spectral analysis to derive peptide ion mass and spectral pattern information. The MS information is used to identify the protein by comparing it with information in a protein database (Shevenko *et al.* (1996) Proc Natl Acad Sci 93:14440-14445). Proteins are separated by

2DE employing isoelectric focusing (IEF) in the first dimension followed by SDS-PAGE in the second dimension. For IEF, an immobilized pH gradient strip is useful to increase reproducibility and resolution of the separation. Alternative techniques may be used to improve resolution of very basic, hydrophobic, or high molecular weight proteins. The separated proteins are detected using a stain or dye such as silver  
5 stain, Coomassie blue, or spiro red (Molecular Probes, Eugene OR) that is compatible with MS. Gels may be blotted onto a PVDF membrane for western analysis and optically scanned using a STORM scanner (APB) to produce a computer-readable output which is analyzed by pattern recognition software such as MELANIE (GeneBio, Geneva, Switzerland). The software annotates individual spots by assigning a unique identifier and calculating their respective x,y coordinates, molecular masses, isoelectric points, and  
10 signal intensity. Individual spots of interest, such as those representing differentially expressed proteins, are excised and proteolytically digested with a site-specific protease such as trypsin or chymotrypsin, singly or in combination, to generate a set of small peptides, preferably in the range of 1-2 kDa. Prior to digestion, samples may be treated with reducing and alkylating agents, and following digestion, the peptides are then separated by liquid chromatography or capillary electrophoresis and analyzed using MS.

15 MS converts components of a sample into gaseous ions, separates the ions based on their mass-to-charge ratio, and determines relative abundance. For peptide mass fingerprinting analysis, a MALDI-TOF (Matrix Assisted Laser Desorption/Ionization-Time of Flight), ESI (Electrospray Ionization), and TOF-TOF (Time of Flight/Time of Flight) machines are used to determine a set of highly accurate peptide masses. Using analytical programs, such as TURBOSEQUENT software (Finnigan, San  
20 Jose CA), the MS data is compared against a database of theoretical MS data derived from known or predicted proteins. A minimum match of three peptide masses is used for reliable protein identification. If additional information is needed for identification, Tandem-MS may be used to derive information about individual peptides. In tandem-MS, a first stage of MS is performed to determine individual peptide masses. Then selected peptide ions are subjected to fragmentation using a technique such as collision  
25 induced dissociation (CID) to produce an ion series. The resulting fragmentation ions are analyzed in a second round of MS, and their spectral pattern may be used to determine a short stretch of amino acid sequence (Dancik et al. (1999) J Comput Biol 6:327-342). Assuming the protein is represented in the database, a combination of peptide mass and fragmentation data, together with the calculated MW and pI of the protein, will usually yield an unambiguous identification. If no match is found, protein sequence  
30 can be obtained using direct chemical sequencing procedures well known in the art (cf. Creighton (1984) Proteins, Structures and Molecular Properties, WH Freeman, New York NY).

#### Chemical Synthesis of Peptides

Proteins or portions thereof may be produced not only by recombinant methods, but also by using chemical methods well known in the art. Solid phase peptide synthesis may be carried out in a batchwise

or continuous flow process which sequentially adds  $\alpha$ -amino- and side chain-protected amino acid residues to an insoluble polymeric support via a linker group. A linker group such as methylamine-derivatized polyethylene glycol is attached to poly(styrene-co-divinylbenzene) to form the support resin. The amino acid residues are N- $\alpha$ -protected by acid labile Boc (t-butyloxycarbonyl) or base-labile Fmoc (9-fluorenylmethoxycarbonyl). The carboxyl group of the protected amino acid is coupled to the amine of the linker group to anchor the residue to the solid phase support resin. Trifluoroacetic acid or piperidine are used to remove the protecting group in the case of Boc or Fmoc, respectively. Each additional amino acid is added to the anchored residue using a coupling agent or pre-activated amino acid derivative, and the resin is washed. The full length peptide is synthesized by sequential deprotection, coupling of derivitized amino acids, and washing with dichloromethane and/or N, N-dimethylformamide. The peptide is cleaved between the peptide carboxy terminus and the linker group to yield a peptide acid or amide. (Novabiochem 1997/98 Catalog and Peptide Synthesis Handbook, San Diego CA pp. S1-S20). Automated synthesis may also be carried out on machines such as the 431A peptide synthesizer (ABI). A protein or portion thereof may be purified by preparative high performance liquid chromatography and its composition confirmed by amino acid analysis or by sequencing (Creighton (1984) Proteins, Structures and Molecular Properties, WH Freeman, New York NY).

#### Antibodies

Antibodies, or immunoglobulins (Ig), are components of immune response expressed on the surface of or secreted into the circulation by B cells. The prototypical antibody is a tetramer composed of two identical heavy polypeptide chains (H-chains) and two identical light polypeptide chains (L-chains) interlinked by disulfide bonds which binds and neutralizes foreign antigens. Based on their H-chain, antibodies are classified as IgA, IgD, IgE, IgG or IgM. The most common class, IgG, is tetrameric while other classes are variants or multimers of the basic structure.

Antibodies are described in terms of their two functional domains. Antigen recognition is mediated by the Fab (antigen binding fragment) region of the antibody, while effector functions are mediated by the Fc (crystallizable fragment) region. The binding of antibody to antigen triggers destruction of the antigen by phagocytic white blood cells such as macrophages and neutrophils. These cells express surface Fc receptors that specifically bind to the Fc region of the antibody and allow the phagocytic cells to destroy antibody-bound antigen. Fc receptors are single-pass transmembrane glycoproteins containing about 350 amino acids whose extracellular portion typically contains two or three Ig domains (Sears et al. (1990) J Immunol 144:371-378).

#### Preparation and Screening of Antibodies

Various hosts including mice, rats, rabbits, goats, llamas, camels, and human cell lines may be immunized by injection with an antigenic determinant. Adjuvants such as Freund's, mineral gels, and

surface active substances such as lysolecithin, pluronic polyols, polyanions, peptides, oil emulsions, keyhole limpet hemacyanin (KLH; Sigma-Aldrich), and dinitrophenol may be used to increase immunological response. In humans, BCG (bacilli Calmette-Guerin) and Corynebacterium parvum increase response. The antigenic determinant may be an oligopeptide, peptide, or protein. When the amount of antigenic determinant allows immunization to be repeated, specific polyclonal antibody with high affinity can be obtained (Klinman and Press (1975) Transplant Rev 24:41-83). Oligopeptides which may contain between about five and about fifteen amino acids identical to a portion of the endogenous protein may be fused with proteins such as KLH in order to produce antibodies to the chimeric molecule.

Monoclonal antibodies may be prepared using any technique which provides for the production of antibodies by continuous cell lines in culture. These include the hybridoma technique, the human B-cell hybridoma technique, and the EBV-hybridoma technique (Kohler et al. (1975) Nature 256:495-497; Kozbor et al. (1985) J Immunol Methods 81:31-42; Cote et al. (1983) Proc Natl Acad Sci 80:2026-2030; and Cole et al. (1984) Mol Cell Biol 62:109-120).

Chimeric antibodies may be produced by techniques such as splicing of mouse antibody genes to human antibody genes to obtain a molecule with appropriate antigen specificity and biological activity (Morrison et al. (1984) Proc Natl Acad Sci 81:6851-6855; Neuberger et al. (1984) Nature 312:604-608; and Takeda et al. (1985) Nature 314:452-454). Alternatively, techniques described for antibody production may be adapted, using methods known in the art, to produce specific, single chain antibodies. Antibodies with related specificity, but of distinct idiotypic composition, may be generated by chain shuffling from random combinatorial immunoglobulin libraries (Burton (1991) Proc Natl Acad Sci 88:10134-10137). Antibody fragments which contain specific binding sites for an antigenic determinant may also be produced. For example, such fragments include, but are not limited to, F(ab')<sub>2</sub> fragments produced by pepsin digestion of the antibody molecule and Fab fragments generated by reducing the disulfide bridges of the F(ab')<sub>2</sub> fragments. Alternatively, Fab expression libraries may be constructed to allow rapid and easy identification of monoclonal Fab fragments with the desired specificity (Huse et al. (1989) Science 246:1275-1281).

Antibodies may also be produced by inducing production in the lymphocyte population or by screening immunoglobulin libraries or panels of highly specific binding reagents as disclosed in Orlandi et al. (1989; Proc Natl Acad Sci 86:3833-3837) or Winter et al. (1991; Nature 349:293-299). A protein may be used in screening assays of phagemid or B-lymphocyte immunoglobulin libraries to identify antibodies having a desired specificity. Numerous protocols for competitive binding or immunoassays using either polyclonal or monoclonal antibodies with established specificities are well known in the art.

#### Antibody Specificity

Various methods such as Scatchard analysis combined with radioimmunoassay techniques may be used to assess the affinity of particular antibodies for a protein. Affinity is expressed as an association constant,  $K_a$ , which is defined as the molar concentration of protein-antibody complex divided by the molar concentrations of free antigen and free antibody under equilibrium conditions. The  $K_a$  determined for a preparation of polyclonal antibodies, which are heterogeneous in their affinities for multiple antigenic determinants, represents the average affinity, or avidity, of the antibodies. The  $K_a$  determined for a preparation of monoclonal antibodies, which are specific for a particular antigenic determinant, represents a true measure of affinity. High-affinity antibody preparations with  $K_a$  ranging from about  $10^9$  to  $10^{12}$  L/mole are commonly used in immunoassays in which the protein-antibody complex must withstand rigorous manipulations. Low-affinity antibody preparations with  $K_a$  ranging from about  $10^6$  to  $10^7$  L/mole are preferred for use in immunopurification and similar procedures which ultimately require dissociation of the protein, preferably in active form, from the antibody (Catty (1988) Antibodies, Volume I: A Practical Approach, IRL Press, Washington DC; Liddell and Cryer (1991) A Practical Guide to Monoclonal Antibodies, John Wiley & Sons, New York NY).

The titer and avidity of polyclonal antibody preparations may be further evaluated to determine the quality and suitability of such preparations for certain downstream applications. For example, a polyclonal antibody preparation containing about 5-10 mg specific antibody/ml, is generally employed in procedures requiring precipitation of protein-antibody complexes. Procedures for making antibodies, evaluating antibody specificity, titer, and avidity, and guidelines for antibody quality and usage in various applications, are discussed in Catty (supra) and Ausubel (supra) pp. 11.1-11.31.

#### Cell Transformation Assays

Cell transformation, the conversion of a normal cell to a cancerous cell, is a highly complex and genetically diverse process. However, certain alterations in cell physiology that are associated with this process can be assayed using either in vitro cell-based systems or in vivo animal models. Known alterations include acquired self-sufficiency relative to growth signals, an insensitivity to growth-inhibitory signals, unlimited replicative potential, evasion of apoptosis, sustained angiogenesis, and cellular invasion and metastasis. See Hanahan and Weinberg (2000) Cell 100:57-70. Such assays can be used, for example, to assess the effect of transfecting a cell with a gene such as HG38, on transformation of the cell.

#### 30 DIAGNOSTICS

Differential expression of HG38, as detected using HG38, a polynucleotide encoding HG38, or an antibody that specifically binds HG38, and at least one of the assays below can be used to diagnose a colon or lung cancer.

#### Labeling of Molecules for Assay



A wide variety of reporter molecules and conjugation techniques are known by those skilled in the art and may be used in various nucleic acid, amino acid, and antibody assays. Synthesis of labeled molecules may be achieved using kits such as those supplied by Promega (Madison WI) or APB for incorporation of a labeled nucleotide such as  $^{32}\text{P}$ -dCTP (APB), Cy3-dCTP or Cy5-dCTP (Qiagen-Operon, Alameda CA), or amino acid such as  $^{35}\text{S}$ -methionine (APB). Nucleotides and amino acids may be directly labeled with a variety of substances including fluorescent, chemiluminescent, or chromogenic agents, and the like, by chemical conjugation to amines, thiols and other groups present in the molecules using reagents such as BIODIPY or FITC (Molecular Probes).

#### Nucleic Acid Assays

10 The polynucleotides, fragments, oligonucleotides, complementary RNAs, and peptide nucleic acids (PNA) may be used to detect and quantify differential gene expression for diagnosis of a disorder. Similarly antibodies which specifically bind HG38 may be used to quantitate the protein. Disorders associated with such differential expression include colon and lung cancer. The diagnostic assay may use hybridization or amplification technology to compare gene expression in a biological sample from a patient to standard samples in order to detect differential gene expression. Qualitative or quantitative methods for this comparison are well known in the art.

#### Expression Profiles

An expression profile comprises the expression of a plurality of polynucleotides or protein as measured using standard assays with a sample. The polynucleotides, proteins or antibodies of the invention may be used as elements on a array to produce an expression profile. In one embodiment, the array is used to diagnose or monitor the progression of disease.

For example, the polynucleotide or probe may be labeled by standard methods and added to a biological sample from a patient under conditions for the formation of hybridization complexes. After an incubation period, the sample is washed and the amount of label (or signal) associated with hybridization complexes, is quantified and compared with a standard value. If complex formation in the patient sample is altered in comparison to either a normal or disease standard, then differential expression indicates the presence of a disorder.

In order to provide standards for establishing differential expression, normal and disease expression profiles are established. This is accomplished by combining a sample taken from normal subjects, either animal or human, with a polynucleotide under conditions for hybridization to occur. Standard hybridization complexes may be quantified by comparing the values obtained using normal subjects with values from an experiment in which a known amount of a purified sequence is used. Standard values obtained in this manner may be compared with values obtained from samples from

patients who were diagnosed with a particular condition, disease, or disorder. Deviation from standard values toward those associated with a particular disorder is used to diagnose or stage that disorder.

By analyzing changes in patterns of gene expression, disease can be diagnosed at earlier stages before the patient is symptomatic. The invention can be used to formulate a prognosis and to design a treatment regimen. The invention can also be used to monitor the efficacy of treatment. For treatments with known side effects, the array is employed to improve the treatment regimen. A dosage is established that causes a change in genetic expression patterns indicative of successful treatment. Expression patterns associated with the onset of undesirable side effects are avoided. This approach may be more sensitive and rapid than waiting for the patient to show inadequate improvement, or to manifest side effects, before altering the course of treatment.

In another embodiment, animal models which mimic a human disease can be used to characterize expression profiles associated with a particular condition, disease, or disorder; or treatment of the condition, disease, or disorder. Novel treatment regimens may be tested in these animal models using arrays to establish and then follow expression profiles over time. In addition, arrays may be used with cell cultures or tissues removed from animal models to rapidly screen large numbers of candidate drug molecules, looking for ones that produce an expression profile similar to those of known therapeutic drugs, with the expectation that molecules with the same expression profile will likely have similar therapeutic effects. Thus, the invention provides the means to rapidly determine the molecular mode of action of a drug.

Such assays may also be used to evaluate the efficacy of a particular therapeutic treatment regimen in animal studies or in clinical trials or to monitor the treatment of an individual patient. Once the presence of a condition is established and a treatment protocol is initiated, diagnostic assays may be repeated on a regular basis to determine if the level of expression in the patient begins to approximate that which is observed in a normal subject. The results obtained from successive assays may be used to show the efficacy of treatment over a period ranging from several days to years.

#### Protein Assays

Immunological methods for detecting and measuring complex formation as a measure of protein expression using either specific polyclonal or monoclonal antibodies are known in the art. Examples of such techniques include antibody arrays, enzyme-linked immunosorbent assays, fluorescence-activated cell sorting, 2D-PAGE and scintillation counting, protein arrays, radioimmunoassays, and western analysis. Such immunoassays typically involve the measurement of complex formation between the protein and its specific antibody. These assays and their quantitation against purified, labeled standards are well known in the art (Ausubel, supra, unit 10.1-10.6). A two-site, monoclonal-based immunoassay

utilizing antibodies reactive to two non-interfering epitopes is preferred, but a competitive binding assay may be employed (Pound (1998) Immunochemical Protocols, Humana Press, Totowa NJ).

These methods are also useful for diagnosing diseases that show differential protein expression. Normal or standard values for protein expression are established by combining body fluids or cell extracts  
5 taken from a normal mammalian or human subject with specific antibodies to a protein under conditions for complex formation. Standard values for complex formation in normal and diseased tissues are established by various methods, often photometric means. Then complex formation as it is expressed in a subject sample is compared with the standard values. Deviation from the normal standard and toward the diseased standard provides parameters for disease diagnosis or prognosis while deviation away from the  
10 diseased and toward the normal standard may be used to evaluate treatment efficacy.

Recently, antibody arrays have allowed the development of techniques for high-throughput screening of recombinant antibodies. Such methods use robots to pick and grid bacteria containing antibody genes, and a filter-based ELISA to screen and identify clones that express antibody fragments. Because liquid handling is eliminated and the clones are arrayed from master stocks, the same antibodies  
15 can be spotted multiple times and screened against multiple antigens simultaneously. Antibody arrays are highly useful in the identification of differentially expressed proteins. (See de Wildt *et al.* (2000) *Nature Biotechnol* 18:989-94.)

#### THERAPEUTICS

Differential expression of polynucleotides encoding HG38 is highly associated with colon  
20 and lung cancer as shown in data presented in Figures 3-8, Tables 1-2, and Northern analysis. HG38 clearly plays a role in colon and lung cancer.

In one embodiment, when decreased expression or activity of the protein is desired, an antibody , antagonist, inhibitor, a pharmaceutical agent or a composition containing one or more of these molecules may be delivered to a subject in need of such treatment. Such delivery may be effected by methods well  
25 known in the art and may include delivery by an antibody that specifically binds the protein. For therapeutic use, monoclonal antibodies are used to block an active site, inhibit dimer formation, trigger apoptosis and the like.

In another embodiment, when increased expression or activity of the protein is desired, the protein, an agonist, an enhancer, a pharmaceutical agent or a composition containing one or more of these  
30 molecules may be delivered to a subject in need of such treatment. Such delivery may be effected by methods well known in the art and may include delivery of a pharmaceutical agent by an antibody specifically targeted to the protein.

Any of the polynucleotides, complementary molecules, or fragments thereof, proteins or portions thereof, vectors delivering these nucleic acid molecules or expressing the proteins, therapeutic antibodies,

and ligands binding the polynucleotide or protein may be administered in combination with other therapeutic agents. Selection of the agents for use in combination therapy may be made by one of ordinary skill in the art according to conventional pharmaceutical principles. A combination of therapeutic agents may act synergistically to affect treatment of a particular disorder at a lower dosage of each agent.

#### 5 Modification of Gene Expression Using Nucleic Acids

Gene expression may be modified by designing complementary or antisense molecules (DNA, RNA, or PNA) to the control, 5', 3', or other regulatory regions of the gene encoding HG38.

Oligonucleotides designed to inhibit transcription initiation are preferred. Similarly, inhibition can be achieved using triple helix base-pairing which inhibits the binding of polymerases, transcription factors, or  
10 regulatory molecules (Gee *et al.* In: Huber and Carr (1994) Molecular and Immunologic Approaches, Futura Publishing, Mt. Kisco NY, pp. 163-177). A complementary molecule may also be designed to block translation by preventing binding between ribosomes and mRNA. In one alternative, a library or plurality of polynucleotides may be screened to identify those which specifically bind a regulatory, nontranslated sequence.

15 Ribozymes, enzymatic RNA molecules, may also be used to catalyze the specific cleavage of RNA. The mechanism of ribozyme action involves sequence-specific hybridization of the ribozyme molecule to complementary target RNA followed by endonucleolytic cleavage at sites such as GUA, GUU, and GUC. Once such sites are identified, an oligonucleotide with the same sequence may be evaluated for secondary structural features which would render the oligonucleotide inoperable. The  
20 suitability of candidate targets may also be evaluated by testing their hybridization with complementary oligonucleotides using ribonuclease protection assays.

Complementary nucleic acids and ribozymes of the invention may be prepared via recombinant expression, *in vitro* or *in vivo*, or using solid phase phosphoramidite chemical synthesis. In addition, RNA molecules may be modified to increase intracellular stability and half-life by addition of flanking  
25 sequences at the 5' and/or 3' ends of the molecule or by the use of phosphorothioate or 2' O-methyl rather than phosphodiesterase linkages within the backbone of the molecule. Modification is inherent in the production of PNAs and can be extended to other nucleic acid molecules. Either the inclusion of nontraditional bases such as inosine, queosine, and wybutosine, or the modification of adenine, cytidine, guanine, thymine, and uridine with acetyl-, methyl-, thio- groups renders the molecule more resistant to  
30 endogenous endonucleases.

#### cDNA Therapeutics

The cDNAs of the invention can be used in gene therapy. cDNAs can be delivered *ex vivo* to target cells, such as cells of bone marrow. Once stable integration and transcription and or translation are confirmed, the bone marrow may be reintroduced into the subject. Expression of the protein encoded by

the cDNA may correct a disorder associated with mutation of a normal sequence, reduction or loss of an endogenous target protein, or overexpression of an endogenous or mutant protein. Alternatively, cDNAs may be delivered in vivo using vectors such as retrovirus, adenovirus, adeno-associated virus, herpes simplex virus, and bacterial plasmids. Non-viral methods of gene delivery include cationic liposomes, polylysine conjugates, artificial viral envelopes, and direct injection of DNA (Anderson (1998) Nature 392:25-30; Dachs et al. (1997) Oncol Res 9:313-325; Chu et al. (1998) J Mol Med 76(3-4):184-192; Weiss et al. (1999) Cell Mol Life Sci 55(3):334-358; Agrawal (1996) Antisense Therapeutics, Humana Press, Totowa NJ; and August et al. (1997) Gene Therapy (Advances in Pharmacology, Vol. 40), Academic Press, San Diego CA).

10

### Monoclonal Antibody Therapeutics

Antibodies, and in particular monoclonal antibodies, that specifically bind a particular protein, enzyme, or receptor and block its overexpression are now being used therapeutically. The first widely accepted therapeutic antibodies were HERCEPTIN (Trastuzumab, Genentech, S. San Francisco CA) and

5 GLEEVEC (imatinib mesylate, Novartis Pharmaceuticals, East Hanover NJ). HERCEPTIN is a humanized antibody approved for the treatment of HER2 positive metastatic breast cancer. It is designed to bind and block the function of overexpressed HER2 protein. GLEEVEC is indicated for the treatment of patients with Philadelphia chromosome positive (Ph+) chronic myeloid leukemia (CML) in blast crisis, accelerated phase, or in chronic phase after failure of interferon-alpha therapy. A second indication for

10 GLEEVEC is treatment of patients with KIT (CD117) positive unresectable and/or metastatic malignant gastrointestinal stromal tumors. Other monoclonal antibodies are in various stages of clinical trials for indications such as prostate cancer, lymphoma, melanoma, pneumococcal infections, rheumatoid arthritis, psoriasis, systemic lupus erythematosus, and the like.

### **Screening and Purification Assays**

15 The polynucleotide encoding HG38 may be used to screen a library or a plurality of molecules or compounds for specific binding affinity. The libraries may be antisense molecules, artificial chromosome constructions, branched nucleic acid molecules, DNA molecules, peptides, peptide nucleic acid, proteins such as transcription factors, enhancers, or repressors, RNA molecules, ribozymes, and other ligands which regulate the activity, replication, transcription, or translation of the endogenous gene. The assay

20 involves combining a polynucleotide with a library or plurality of molecules or compounds under conditions allowing specific binding, and detecting specific binding to identify at least one molecule which specifically binds the polynucleotide.

The polynucleotide of the invention may be incubated with a plurality of purified molecules or compounds and binding activity determined by methods well known in the art, e.g., a gel-retardation assay

25 (USPN 6,010,849) or a reticulocyte lysate transcriptional assay. The polynucleotide may be incubated with nuclear extracts from biopsied and/or cultured cells and tissues. Specific binding between the polynucleotide and a molecule or compound in the nuclear extract is initially determined by gel shift assay and may be later confirmed by recovering and raising antibodies against that molecule or compound. When these antibodies are added into the assay, they cause a supershift in the gel-retardation assay.

30 The polynucleotide may be used to purify a molecule or compound using affinity chromatography methods well known in the art. In one embodiment, the polynucleotide is chemically reacted with cyanogen bromide groups on a polymeric resin or gel. Then a sample is passed over and reacts with or binds to the polynucleotide. The molecule or compound which is bound to the polynucleotide may be

released from the polynucleotide by increasing the salt concentration of the flow-through medium and collected.

The protein or a portion thereof may be used to purify a ligand from a sample. A method for using a protein to purify a ligand would involve combining the protein with a sample under conditions to allow specific binding, detecting specific binding between the protein and ligand, recovering the bound protein, and using a chaotropic agent to separate the protein from the purified ligand.

HG38 may be used to screen a plurality of molecules or compounds in any of a variety of screening assays. The portion of the protein employed in such screening may be free in solution, affixed to an abiotic or biotic substrate (e.g. borne on a cell surface), or located intracellularly. For example, in one method, viable or fixed prokaryotic host cells that are stably transformed with recombinant nucleic acids that have expressed and positioned a peptide on their cell surface can be used in screening assays. The cells are screened against a plurality or libraries of ligands, and the specificity of binding or formation of complexes between the expressed protein and the ligand can be measured. Depending on the particular kind of molecules or compounds being screened, the assay may be used to identify agonists, antagonists, antibodies, DNA molecules, small drug molecules, immunoglobulins, inhibitors, mimetics, peptides, peptide nucleic acids, proteins, and RNA molecules or any other ligand, which specifically binds the protein.

In one aspect, this invention contemplates a method for high throughput screening using very small assay volumes and very small amounts of test compound as described in USPN 5,876,946, incorporated herein by reference. This method is used to screen large numbers of molecules and compounds via specific binding. In another aspect, this invention also contemplates the use of competitive drug screening assays in which neutralizing antibodies capable of binding the protein specifically compete with a test compound capable of binding to the protein. Molecules or compounds identified by screening may be used in a mammalian model system to evaluate their toxicity or therapeutic potential.

## 25 Pharmaceutical Compositions

Pharmaceutical compositions may be formulated and administered, to a subject in need of such treatment, to attain a therapeutic effect. Such compositions contain the instant protein, agonists, antagonists, bispecific molecules, small drug molecules, immunoglobulins, inhibitors, mimetics, multispecific molecules, peptides, peptide nucleic acids, pharmaceutical agent, proteins, and RNA molecules. Compositions may be manufactured by conventional means such as mixing, dissolving, granulating, dragee-making, levigating, emulsifying, encapsulating, entrapping, or lyophilizing. The composition may be provided as a salt, formed with acids such as hydrochloric, sulfuric, acetic, lactic, tartaric, malic, and succinic, or as a lyophilized powder which may be combined with a sterile buffer such

as saline, dextrose, or water. These compositions may include auxiliaries or excipients which facilitate processing of the active compounds.

Auxiliaries and excipients may include coatings, fillers or binders including sugars such as lactose, sucrose, mannitol, glycerol, or sorbitol; starches from corn, wheat, rice, or potato; proteins such as albumin, gelatin and collagen; cellulose in the form of hydroxypropylmethyl-cellulose, methyl cellulose, or sodium carboxymethylcellulose; gums including arabic and tragacanth; lubricants such as magnesium stearate or talc; disintegrating or solubilizing agents such as the, agar, alginic acid, sodium alginate or cross-linked polyvinyl pyrrolidone; stabilizers such as carbopol gel, polyethylene glycol, or titanium dioxide; and dyestuffs or pigments added for identify the product or to characterize the quantity of active compound or dosage.

These compositions may be administered by any number of routes including oral, intravenous, intramuscular, intra-arterial, intramedullary, intrathecal, intraventricular, transdermal, subcutaneous, intraperitoneal, intranasal, enteral, topical, sublingual, or rectal.

The route of administration and dosage will determine formulation; for example, oral administration may be accomplished using tablets, pills, dragees, capsules, liquids, gels, syrups, slurries, or suspensions; parenteral administration may be formulated in aqueous, physiologically compatible buffers such as Hanks' solution, Ringer's solution, or physiologically buffered saline. Suspensions for injection may be aqueous, containing viscous additives such as sodium carboxymethyl cellulose or dextran to increase the viscosity, or oily, containing lipophilic solvents such as sesame oil or synthetic fatty acid esters such as ethyl oleate or triglycerides, or liposomes. Penetrants well known in the art are used for topical or nasal administration.

#### Toxicity and Therapeutic Efficacy

A therapeutically effective dose refers to the amount of active ingredient which ameliorates symptoms or condition. For any compound, a therapeutically effective dose can be estimated from cell culture assays using normal and neoplastic cells or in animal models. Therapeutic efficacy, toxicity, concentration range, and route of administration may be determined by standard pharmaceutical procedures using experimental animals.

The therapeutic index is the dose ratio between therapeutic and toxic effects--LD50 (the dose lethal to 50% of the population)/ED50 (the dose therapeutically effective in 50% of the population)--and large therapeutic indices are preferred. Dosage is within a range of circulating concentrations, includes an ED50 with little or no toxicity, and varies depending upon the composition, method of delivery, sensitivity of the patient, and route of administration. Exact dosage will be determined by the practitioner in light of factors related to the subject in need of the treatment.



Dosage and administration are adjusted to provide active moiety that maintains therapeutic effect. Factors for adjustment include the severity of the disease state, general health of the subject, age, weight, and gender of the subject, diet, time and frequency of administration, drug combination(s), reaction sensitivities, and tolerance/response to therapy. Long-acting pharmaceutical compositions may be administered every 3 to 4 days, every week, or once every two weeks depending on half-life and clearance rate of the particular composition.

Normal dosage amounts may vary from 0.1  $\mu\text{g}$ , up to a total dose of about 1 g, depending upon the route of administration. The dosage of a particular composition may be lower when administered to a patient in combination with other agents, drugs, or hormones. Guidance as to particular dosages and methods of delivery is provided in the pharmaceutical literature. Further details on techniques for formulation and administration may be found in the latest edition of Remington's Pharmaceutical Sciences (Mack Publishing, Easton PA).

#### Model Systems

Animal models may be used as bioassays where they exhibit a phenotypic response similar to that of humans and where exposure conditions are relevant to human exposures. Mammals are the most common models, and most infectious agent, cancer, drug, and toxicity studies are performed on rodents such as rats or mice because of low cost, availability, lifespan, gestation period, numbers of progeny, and abundant reference literature. Inbred and outbred rodent strains provide a convenient model for investigation of the physiological consequences of under- or over-expression of genes of interest and for the development of methods for diagnosis and treatment of diseases. A mammal inbred to over-express a particular gene (for example, secreted in milk) may also serve as a convenient source of the protein expressed by that gene.

#### Toxicology

Toxicology is the study of the effects of agents on living systems. The majority of toxicity studies are performed on rats or mice. Observation of qualitative and quantitative changes in physiology, behavior, homeostatic processes, and lethality in the rats or mice are used to generate a toxicity profile and to assess consequences on human health following exposure to the agent.

Genetic toxicology identifies and analyzes the effect of an agent on the rate of endogenous, spontaneous, and induced genetic mutations. Genotoxic agents usually have common chemical or physical properties that facilitate interaction with nucleic acids and are most harmful when chromosomal aberrations are transmitted to progeny. Toxicological studies may identify agents that increase the frequency of structural or functional abnormalities in the tissues of the progeny if administered to either parent before conception, to the mother during pregnancy, or to the developing organism. Mice and rats

are most frequently used in these tests because their short reproductive cycle allows the production of the numbers of organisms needed to satisfy statistical requirements.

Acute toxicity tests are based on a single administration of an agent to the subject to determine the symptomology or lethality of the agent. Three experiments are conducted: 1) an initial dose-range-finding experiment, 2) an experiment to narrow the range of effective doses, and 3) a final experiment for establishing the dose-response curve.

Subchronic toxicity tests are based on the repeated administration of an agent. Rat and dog are commonly used in these studies to provide data from species in different families. With the exception of carcinogenesis, there is considerable evidence that daily administration of an agent at high-dose concentrations for periods of three to four months will reveal most forms of toxicity in adult animals.

Chronic toxicity tests, with a duration of a year or more, are used to test whether long term administration may elicit toxicity, teratogenesis, or carcinogenesis. When studies are conducted on rats, a minimum of three test groups plus one control group are used, and animals are examined and monitored at the outset and at intervals throughout the experiment.

#### 15 Transgenic Animal Models

Transgenic rodents that over-express or under-express a gene of interest may be inbred and used to model human diseases or to test therapeutic or toxic agents. (See, e.g., USPN 5,175,383 and USPN 5,767,337.) In some cases, the introduced gene may be activated at a specific time in a specific tissue type during fetal or postnatal development. Expression of the transgene is monitored by analysis of phenotype, of tissue-specific mRNA expression, or of serum and tissue protein levels in transgenic animals before, during, and after challenge with experimental drug therapies.

#### Embryonic Stem Cells

Embryonic (ES) stem cells isolated from rodent embryos retain the ability to form embryonic tissues. When ES cells are placed inside a carrier embryo, they resume normal development and contribute to tissues of the live-born animal. ES cells are the preferred cells used in the creation of experimental knockout and knockin rodent strains. Mouse ES cells, such as the mouse 129/SvJ cell line, are derived from the early mouse embryo and are grown under culture conditions well known in the art. Vectors used to produce a transgenic strain contain a disease gene candidate and a marker gene, the latter serves to identify the presence of the introduced disease gene. The vector is transformed into ES cells by methods well known in the art, and transformed ES cells are identified and microinjected into mouse cell blastocysts such as those from the C57BL/6 mouse strain. The blastocysts are surgically transferred to pseudopregnant dams, and the resulting chimeric progeny are genotyped and bred to produce heterozygous or homozygous strains.

ES cells derived from human blastocysts may be manipulated in vitro to differentiate into at least eight separate cell lineages. These lineages are used to study the differentiation of various cell types and tissues in vitro, and they include endoderm, mesoderm, and ectodermal cell types which differentiate into, for example, neural cells, hematopoietic lineages, and cardiomyocytes.

#### 5 Knockout Analysis

In gene knockout analysis, a region of a gene is enzymatically modified to include a non-mammalian gene such as the neomycin phosphotransferase gene (neo; Capecchi (1989) Science 244:1288-1292). The modified gene is transformed into cultured ES cells and integrates into the endogenous genome by homologous recombination. The inserted sequence disrupts transcription and translation of the  
10 endogenous gene. Transformed cells are injected into rodent blastulae, and the blastulae are implanted into pseudopregnant dams. Transgenic progeny are crossbred to obtain homozygous inbred lines which lack a functional copy of the mammalian gene. In one example, the mammalian gene is a human gene.

#### Knockin Analysis

ES cells can be used to create knockin humanized animals (pigs) or transgenic animal models  
15 (mice or rats) of human diseases. With knockin technology, a region of a human gene is injected into animal ES cells, and the human sequence integrates into the animal cell genome. Transformed cells are injected into blastulae and the blastulae are implanted as described above. Transgenic progeny or inbred lines are studied and treated with pharmaceutical agents to obtain information on treatment of the analogous human condition. These methods have been used to model several human diseases.

#### 20 Non-Human Primate Model

The field of animal testing deals with data and methodology from basic sciences such as physiology, genetics, chemistry, pharmacology and statistics. These data are paramount in evaluating the effects of therapeutic agents on non-human primates as they can be related to human health. Monkeys are used as human surrogates in vaccine and drug evaluations, and their responses are relevant to human  
25 exposures under similar conditions. Cynomolgus and Rhesus monkeys (Macaca fascicularis and Macaca mulatta, respectively) and Common Marmosets (Callithrix jacchus) are the most common non-human primates (NHPs) used in these investigations. Since great cost is associated with developing and maintaining a colony of NHPs, early research and toxicological studies are usually carried out in rodent models. In studies using behavioral measures such as drug addiction, NHPs are the first choice test  
30 animal. In addition, NHPs and individual humans exhibit differential sensitivities to many drugs and toxins and can be classified as a range of phenotypes from "extensive metabolizers" to "poor metabolizers" of these agents.

In additional embodiments, the polynucleotides which encode the protein may be used in any molecular biology techniques that have yet to be developed, provided the new techniques rely on

properties of polynucleotides that are currently known, including, but not limited to, such properties as the triplet genetic code and specific base pair interactions.

## EXAMPLES

### I cDNA Library Construction

5 Cells or tissues were homogenized and lysed in guanidinium isothiocyanate, in phenol or in a suitable mixture of denaturants such as TRIZOL reagent (Invitrogen) and guanidine isothiocyanate. The lysates were centrifuged over CsCl cushions or extracted with chloroform. RNA was precipitated from the lysates with either isopropanol or sodium acetate and ethanol or by other routine methods.

Phenol extraction and precipitation of RNA were repeated as necessary to increase RNA purity.

10 In some cases, RNA was treated with DNase. For most libraries, poly(A)+ RNA was isolated using oligo d(T)-coupled paramagnetic particles (Promega), OLIGOTEX latex particles (Qiagen, Chatsworth CA), or an OLIGOTEX mRNA purification kit (Qiagen). Alternatively, RNA was isolated directly from tissue lysates using RNA isolation kits such as the POLY(A)PURE mRNA purification kit (Ambion, Austin TX).

In some cases, Stratagene was provided with RNA and constructed the cDNA libraries. cDNA  
15 was synthesized and cDNA libraries were constructed with the UNIZAP vector system (Stratagene) or SUPERScript plasmid system (Invitrogen), using the recommended procedures or similar methods known in the art (Ausubel, *supra*, units 5.1-6.6). Reverse transcription was initiated using oligo d(T) or random primers. Synthetic oligonucleotide adapters were ligated to double stranded cDNA, and the cDNA was digested with appropriate restriction enzymes. For most libraries, the cDNA was size-selected  
20 (300-1000 bp) using SEPHACRYL S1000, SEPHAROSE CL2B, or SEPHAROSE CL4B column chromatography (APB) or preparative agarose gel electrophoresis. cDNAs were ligated into compatible restriction enzyme sites of the polylinker of a suitable plasmid such as pBLUESCRIPT plasmid or pBK-CMV plasmid (both Stratagene), pSPORT1 plasmid or PCDNA2.1 plasmid (both Invitrogen), pINCY (Incyte Genomics, Palo Alto CA), or derivatives thereof. Recombinant plasmids were transformed into  
25 competent *E. coli* cells including XL1-Blue, XL1-BlueMRF, or SOLR (Stratagene) or DH5 $\alpha$ , DH10B, or ElectroMAX DH10B (Invitrogen).

### II Isolation, Preparation, and Sequencing of cDNAs

Plasmids were recovered from host cells by *in vivo* excision using the UNIZAP vector system (Stratagene) or by cell lysis. Plasmids were purified using at least one of the following: a Magic or  
30 WIZARD Minipreps DNA purification system (Promega); an AGTC Miniprep purification kit (Edge Biosystems, Gaithersburg MD); and QIAWELL 8 Plasmid, QIAWELL 8 Plus Plasmid, QIAWELL 8 Ultra Plasmid purification systems or REAL PREP 96 plasmid purification kit from Qiagen. Following precipitation, plasmids were resuspended in 0.1 ml of distilled water and stored, with or without lyophilization, at 4C.

Alternatively, plasmid DNA was amplified from host cell lysates using direct link PCR in a high-throughput format (Rao (1994) Anal Biochem 216:1-14). Host cell lysis and thermal cycling steps were carried out in a single reaction mixture. Samples were processed and stored in 384-well plates, and the concentration of amplified plasmid DNA was quantified fluorometrically using PICOGREEN dye  
 5 (Molecular Probes, Eugene OR) and a FLUOROSKAN II fluorescence scanner (Labsystems Oy, Helsinki, Finland).

Sequencing reactions were processed using standard methods or high-throughput instrumentation such as the CATALYST 800 (ABI) thermal cycler or the DNA ENGINE thermal cycler (MJ Research) in conjunction with the HYDRA microdispenser (Robbins Scientific) or the MICROLAB 2200 (Hamilton)  
 10 liquid transfer system. cDNA sequencing reactions were prepared using reagents obtained from APB or supplied in sequencing kits such as the PRISM BIGDYE Terminator cycle sequencing ready reaction kit (ABI). Electrophoretic separation of cDNA sequencing reactions and detection of labeled polynucleotides were carried out using the MEGABACE 1000 DNA sequencing system (APB) or PRISM 373 or 377 sequencing systems (ABI) in conjunction with standard protocols and base calling software. Reading  
 15 frames within the cDNA sequences were identified using standard methods (Ausubel, supra, unit 7.7).

### III Extension of cDNAs

The cDNAs were extended using the cDNA clone and oligonucleotide primers. One primer was synthesized to initiate 5' extension of the known fragment, and the other, to initiate 3' extension of the known fragment. The initial primers were designed using primer analysis software to be about 22 to 30  
 20 nucleotides in length, to have a GC content of about 50% or more, and to anneal to the target sequence at temperatures of about 68C to about 72C. Any stretch of nucleotides that would result in hairpin structures and primer-primer dimerizations was avoided.

Selected cDNA libraries were used as templates to extend the sequence. If extension was performed than one time, additional or nested sets of primers were designed. Preferred libraries have been  
 25 size-selected to include larger cDNAs and random primed to contain more sequences with 5' or upstream regions of genes. Genomic libraries can be used to obtain regulatory elements extending into the 5' promoter binding region.

High fidelity amplification was obtained by PCR using methods such as that taught in USPN 5,932,451. PCR was performed in 96-well plates using the DNA ENGINE thermal cycler (MJ Research).  
 30 The reaction mix contained DNA template, 200 nmol of each primer, reaction buffer containing  $Mg^{2+}$ ,  $(NH_4)_2SO_4$ , and  $\beta$ -mercaptoethanol, Taq DNA polymerase (APB), ELONGASE enzyme (Invitrogen), and Pfu DNA polymerase (Stratagene), with the following parameters for primer pair PCI A and PCI B (Incyte Genomics): The parameters for the cycles are 1: 94C, three min; 2: 94C, 15 sec; 3: 60C, one min; 4: 68C, two min; 5: 2, 3, and 4 repeated 20 times; 6: 68C, five min; and 7: storage at 4C. In the

alternative, the parameters for primer pair T7 and SK+ (Stratagene) were as follows: 1: 94C, three min; 2: 94C, 15 sec; 3: 57C, one min; 4: 68C, two min; 5: 2, 3, and 4 repeated 20 times; 6: 68C, five min; and 7: storage at 4C.

The concentration of DNA in each well was determined by dispensing 100  $\mu$ l PICOGREEN  
5 quantitation reagent (0.25% reagent in 1x TB, v/v; Molecular Probes) and 0.5  $\mu$ l of undiluted PCR product into each well of an opaque fluorimeter plate (Corning Life Sciences, Acton MA) and allowing the DNA to bind to the reagent. The plate was scanned in a Fluoroskan II (Labsystems Oy, Helsinki Finland) to measure the fluorescence of the sample and to quantify the concentration of DNA. A 5  $\mu$ l to 10  $\mu$ l aliquot of the reaction mixture was analyzed by electrophoresis on a 1% agarose minigel to determine which  
10 reactions were successful in extending the sequence.

The extended clones were desalted, concentrated, transferred to 384-well plates, digested with CviJI cholera virus endonuclease (Molecular Biology Research, Madison WI), and sonicated or sheared prior to religation into pUC18 vector (APB). For shotgun sequences, the digested nucleotide sequences were separated on low concentration (0.6 to 0.8%) agarose gels, fragments were excised, and the agar was  
15 digested with AGARACE enzyme (Promega). Extended clones were religated using T4 DNA ligase (New England Biolabs) into pUC18 vector (APB), treated with Pfu DNA polymerase (Stratagene) to fill-in restriction site overhangs, and transfected into *E. coli* competent cells. Transformed cells were selected on antibiotic-containing media, and individual colonies were picked and cultured overnight at 37C in 384-well plates in LB/2x carbenicillin liquid media.

20 The cells were lysed, and DNA was amplified using primers, Taq DNA polymerase (APB) and Pfu DNA polymerase (Stratagene) with the following parameters: 1: 94C, three min; 2: 94C, 15 sec; 3: 60C, one min; 4: 72C, two min; 5: 2, 3, and 4 repeated 29 times; 6: 72C, five min; and 7: storage at 4C. DNA was quantified using PICOGREEN quantitation reagent (Molecular Probes) as described above. Samples with low DNA recoveries were reamplified using the conditions described above. Samples were  
25 diluted with 20% dimethylsulfoxide (DMSO; 1:2, v/v), and sequenced using DYENAMIC energy transfer sequencing primers and the DYENAMIC DIRECT cycle sequencing kit (APB) or the PRISM BIGDYE terminator cycle sequencing kit (ABI).

#### IV Homology Searching of cDNA Clones and Their Deduced Proteins

The polynucleotides of the Sequence Listing or their deduced amino acid sequences were used to  
30 query databases such as GenBank, SwissProt, BLOCKS, and the like. These databases that contain previously identified and annotated sequences or domains were searched using BLAST or BLAST2 to produce alignments and to determine which sequences were exact matches or homologs. The alignments were to sequences of prokaryotic (bacterial) or eukaryotic (animal, fungal, or plant) origin. Alternatively, algorithms such as the one described in Smith and Smith (1992, Protein Engineering 5:35-51) could have

been used to deal with primary sequence patterns and secondary structure gap penalties. All of the sequences disclosed in this application have lengths of at least 49 nucleotides, and no more than 12% uncalled bases (where N is recorded rather than A, C, G, or T).

As detailed in Karlin and Altschul (1993; Proc Natl Acad Sci 90:5873-5877), BLAST matches  
5 between a query sequence and a database sequence were evaluated statistically and only reported when they satisfied the threshold of  $10^{-25}$  for nucleotides and  $10^{-14}$  for peptides. Homology was also evaluated by product score calculated as follows: the % nucleotide or amino acid identity [between the query and reference sequences] in BLAST is multiplied by the % maximum possible BLAST score [based on the lengths of query and reference sequences] and then divided by 100. In comparison with hybridization  
10 procedures used in the laboratory, the stringency for an exact match was set from a lower limit of about 40 (with 1-2% error due to uncalled bases) to a 100% match of about 70.

The BLAST software suite (NCBI, Bethesda MD), includes various sequence analysis programs including "blastn" that is used to align nucleotide sequences and BLAST2 that is used for direct pairwise comparison of either nucleotide or amino acid sequences. BLAST programs are commonly used with gap  
15 and other parameters set to default settings, e.g.: Matrix: BLOSUM62; Reward for match: 1; Penalty for mismatch: -2; Open Gap: 5 and Extension Gap: 2 penalties; Gap x drop-off: 50; Expect: 10; Word Size: 11; and Filter: on. Identity is measured over the entire length of a sequence. Brenner (*supra*) analyzed BLAST for its ability to identify structural homologs by sequence identity and found 30% identity is a reliable threshold for sequence alignments of at least 150 residues and 40%, for alignments of at least 70  
20 residues.

The polynucleotides of this application were compared with assembled consensus sequences or templates found in the LIFESEQ GOLD database (Incyte Genomics). Component sequences from polynucleotide, extension, full length, and shotgun sequencing projects were subjected to PHRED analysis and assigned a quality score. All sequences with an acceptable quality score were subjected to various  
25 pre-processing and editing pathways to remove low quality 3' ends, vector and linker sequences, polyA tails, Alu repeats, mitochondrial and ribosomal sequences, and bacterial contamination sequences. Edited sequences had to be at least 50 bp in length, and low-information sequences and repetitive elements such as dinucleotide repeats, Alu repeats, and the like, were replaced by "Ns" or masked.

Edited sequences were subjected to assembly procedures in which the sequences were assigned to  
30 gene bins. Each sequence could only belong to one bin, and sequences in each bin were assembled to produce a template. Newly sequenced components were added to existing bins using BLAST and CROSSMATCH. To be added to a bin, the component sequences had to have a BLAST quality score greater than or equal to 150 and an alignment of at least 82% local identity. The sequences in each bin were assembled using PHRAP. Bins with several overlapping component sequences were assembled

using DEEP PHRAP. The orientation of each template was determined based on the number and orientation of its component sequences.

Bins were compared to one another, and those having local similarity of at least 82% were combined and reassembled. Bins having templates with less than 95% local identity were split.

- 5 Templates were subjected to analysis by STITCHER/EXON MAPPER algorithms that determine the probabilities of the presence of splice variants, alternatively spliced exons, splice junctions, differential expression of alternative spliced genes across tissue types or disease states, and the like. Assembly procedures were repeated periodically, and templates were annotated using BLAST against GenBank databases such as GBpri. An exact match was defined as having from 95% local identity over 200 base  
10 pairs through 100% local identity over 100 base pairs and a homology match as having an E-value (or probability score) of  $\leq 1 \times 10^{-8}$ . The templates were also subjected to frameshift FASTx against GENPEPT, and homology match was defined as having an E-value of  $\leq 1 \times 10^{-8}$ . Template analysis and assembly was described in USSN 09/276,534, filed March 25, 1999.

- Following assembly, templates were subjected to BLAST, motif, and other functional analyses  
15 and categorized in protein hierarchies using methods described in USSN 08/812,290 and USSN 08/811,758, both filed March 6, 1997; in USSN 08/947,845, filed October 9, 1997; and in USSN 09/034,807, filed March 4, 1998. Then templates were analyzed by translating each template in all three forward reading frames and searching each translation against the PFAM database of hidden Markov model-based protein families and domains using the HMMER software package (Washington University  
20 School of Medicine, St. Louis MO). The polynucleotide was further analyzed using MACDNASIS PRO software (Hitachi Software Engineering), and LASERGENE software (DNASTAR) and queried against public databases such as the GenBank rodent, mammalian, vertebrate, prokaryote, and eukaryote databases, SwissProt, BLOCKS, PRINTS, PFAM, and Prosite.

## V Northern Analysis, Transcript Imaging

### 25 Northern analysis

Northern analysis is a laboratory technique used to detect the presence of a transcript of a gene and involves the hybridization of a labeled nucleotide sequence to a membrane on which RNAs from a particular cell type or tissue have been bound. The technique is described below and in Ausubel, *supra*, units 4.1-4.9 and was used to generate the data presented in the Tables at pages 13 and 14, above.

- 30 Analogous computer techniques applying BLAST are used to search for identical or related molecules in nucleotide databases such as GenBank or the LIFESEQ database (Incyte Genomics). This analysis is faster than multiple membrane-based hybridizations. In addition, the sensitivity of the computer search can be modified to determine whether any particular match is categorized as exact or homologous. The basis of the search is the product score which was described above in EXAMPLE IV.



The results of northern analysis are reported as a list of libraries in which the transcript encoding HG38 occurs. Abundance and percent abundance are also reported. Abundance directly reflects the number of times a particular transcript is represented in a cDNA library, and percent abundance is abundance divided by the total number of sequences examined in the cDNA library.

#### 5 Transcript Imaging

A transcript image was performed using the LIFESEQ GOLD database (Incyte Genomics). This process allows assessment of the relative abundance of the expressed polynucleotides in all of the cDNA libraries and was described in USPN 5,840,484, incorporated herein by reference. All sequences and cDNA libraries in the LIFESEQ database are categorized by system, organ/tissue and cell type. The  
10 categories include cardiovascular system, connective tissue, digestive system, embryonic structures, endocrine system, exocrine glands, female and male genitalia, germ cells, hemic/immune system, liver, musculoskeletal system, nervous system, pancreas, respiratory system, sense organs, skin, stomatognathic system, unclassified/mixed, and the urinary tract. Criteria for transcript imaging are selected from category, number of cDNAs per library, library description, disease indication, clinical relevance of  
15 sample, and the like.

For each category, the number of libraries in which the sequence was expressed are counted and shown over the total number of libraries in that category. For each library, the number of cDNAs are counted and shown over the total number of cDNAs in that library. In some transcript images, all enriched, normalized or subtracted libraries, which have high copy number sequences can be removed  
20 prior to processing, and all mixed or pooled tissues, which are considered non-specific in that they contain more than one tissue type or more than one subject's tissue, can be excluded from the analysis. Treated and untreated cell lines and/or fetal tissue data can also be excluded where clinical relevance is emphasized. Conversely, fetal tissue can be emphasized wherever elucidation of inherited disorders or differentiation of particular adult or embryonic stem cells into tissues or organs (such as heart, kidney,  
25 nerves or pancreas) would be aided by removing clinical samples from the analysis. Transcript imaging can also be used to support data from other methodologies such as hybridization, guilt-by-association and array technologies.

#### VI Chromosome Mapping

Radiation hybrid and genetic mapping data available from public resources such as the Stanford  
30 Human Genome Center (SHGC), Whitehead Institute for Genome Research (WIGR), and Généthon are used to determine if any of the polynucleotides presented in the Sequence Listing have been mapped. Any of the fragments of the polynucleotide encoding HG38 that have been mapped result in the assignment of all related regulatory and coding sequences to the same location. The genetic map locations are described

as ranges, or intervals, of human chromosomes. The map position of an interval, in cM (which is roughly equivalent to 1 megabase of human DNA), is measured relative to the terminus of the chromosomal p-arm.

## VII Hybridization and Amplification Technologies and Analyses

### Tissue Sample Preparation

- 5 Normal and cancerous colon and lung tissue samples are described by donor identification number in the table below. The first column shows the donor ID; the second, donor age/sex; the third column, a description of the disorder, the fourth column, classification of the tumor; and the fifth column, the source.

	<u>Donor</u>	<u>Age/Sex*</u>	<u>Tissue and Description</u>	<u>Stage</u>	<u>Source</u>
	3579	55/M	colon; well differentiated adenoCA	Dukes' C; TMN T2N1	HCI
10	3580	38/M	colon; poorly differentiated, metastatic adenoCA	T3N1MX	HCI
	3581	U/M	rectal; tumor	NA	HCI
	3582	78/M	colon; moderately differentiated adenoCA	TMN T4N2MX	HCI
	3583	58/M	colon; tubulovillous adenoma (hyperplastic polyp)	NA	HCI
	3647	83/U	colon; invasive moderately differentiated adenoCA	TMN T3N1MX	HCI
15	3649	86/U	colon; invasive well-differentiated adenoCA	NA	HCI
	3479	68/M	colon; adenoCA	NA	HCI
	3839	59/M	colon tumor	NA	HCI
	3757	75/F	colon tumor	NA	HCI
	3756	78/U	colon tumor	NA	HCI
20	4614	67/U	colon; moderately differentiated adenoCA	Dukes' B; TMN T3N0	HCI
	9573	60/F	colon; moderately differentiated adenoCA	Dukes C; T2N2M0	Asterand
	9574	34/F	colon; well differentiated metastatic adenoCA	Dukes C; T2N1M0	Asterand
	9575	60/M	colon; moderately differentiated metastatic adenoCA	Dukes C; TXN1-2M0	Asterand
	9576	65/M	colon; well differentiated adenoCA	Dukes C; T3N2M0	Asterand
25	9577	46/F	colon; well differentiated adenoCA	Dukes C; TXN1-2M0	Asterand
	8401	57/M	colon; well-moderately differentiated adenoCA	Grade II	Asterand
	8403	54/F	colon; adenoCA	Grade II	Asterand
	7162	73/M	lung; poorly differentiated, large cell endocrine	IIB	RCI
	7164	79/M	lung; pulmonary carcinoid	stage IA	RCI
30	7168	75/M	lung; poorly differentiated adenoCA	IB	RCI
	7173	70/M	lung; moderately differentiated squamous cell CA	IIB	RCI
	7175	67/M	lung; moderately differentiated adenoCA	IB	RCI
	7176	72/M	lung; poorly differentiated, adenosquamous	IB	RCI
	7178	68/F	moderately differentiated squamous cell carcinoma	IIIA	RCI
35	7186	61/M	lung; atypical carcinoid	stage IA	RCI
	7188	54/M	lung; poorly differentiated adenoCA	IIIA	RCI
	7189	78/M	lung; poorly differentiated adenoCA	IB	RCI
	7190	50/F	lung; moderately differentiated squamous cell CA	IB	RCI
	7191	43/M	lung; poorly differentiated, squamous cell CA	IIB	RCI
40	7963	71/M	lung; poorly differentiated adenoCA	IIIA	RCI
	9751	70F	lung; poorly differentiated, squamous cell CA	NA	RCI
	9752	56/M	lung, squamous cell carcinoma	NA	RCI
	9753	66/M	lung; moderately differentiated squamous cell CA	NA	RCI
	9754	72/M	lung; moderately differentiated squamous cell CA	NA	RCI
45	9757	58/F	lung, adenoCA	NA	RCI
	9758	48/M	lung, adenoCA	NA	RCI

	9764	73/M	lung, adenoCA	NA	RCI
	5793	73/M	lung; moderately differentiated squamous cell CA	IIA	RCI
	5795	71/F	lung; moderately differentiated adenoCA	IA	RCI
	5796	66/M	lung; moderately differentiated squamous cell CA	IB	RCI
5	5797	73/M	lung; moderately differentiated squamous cell CA	IIB	RCI
	5798	66/F	lung; adenoCA	NA	RCI
	5799	66/F	lung; moderately differentiated adenoCA	IIB	RCI
	5800	75/F	lung; moderately differentiated squamous cell CA	IB	RCI

\* Abbreviations: CA=carcinoma, U=unknown, NA=not available

10 In Figure 2, the normalized, first-strand synthesis, polynucleotide preparations of normal, human heart, brain (whole), lung, liver, skeletal muscle, kidney, pancreas, spleen, thymus, prostate, ovary, small intestine, peripheral blood leukocyte, and colon tissues were obtained from Clontech. Additional polynucleotide preparations of human, adult, normal thyroid, pituitary, and adrenal tissues were obtained from Clinomics Bioscience (Pittsfield MA).

15 The colon cell lines shown in Figure 5 were obtained from ATCC and cultured according to the suppliers specifications. The table below describes cancerous and non-cancerous human cell lines analyzed in Figure 5 for HG38 expression. The first column lists the name of the cell line, the second column, the tissue source, the third column, a description of the cell line, and the fourth and fifth columns whether the cell line is tumorigenic and/or metastatic in mice.

20	Sample	Tissue	Description	Tumorigenic	Metastatic
	LS123	Colon	adenocarcinoma	No	No
	Caco-2	Colon	adenocarcinoma	Yes	No
	HCT 116	Colon	adenocarcinoma	Yes	No
	HT-29	Colon	adenocarcinoma	Yes	No
25	LS 174T	Colon	adenocarcinoma	Yes	No
	SW480	Colon	adenocarcinoma	Yes	No
	COLO 205	Colon	ascites metastasis from adenocarcinoma	Yes	Yes
	SW620	Colon	lymph node metastasis from adenocarcinoma (same patient as SW480)	Yes	Yes

30

### Immobilization of polynucleotides on a Substrate

The polynucleotides are applied to a substrate by one of the following methods. A mixture of polynucleotides is fractionated by gel electrophoresis and transferred to a nylon membrane by capillary transfer. Alternatively, the polynucleotides are individually ligated to a vector and inserted into bacterial host cells to form a library. The polynucleotides are then arranged on a substrate by one of the following methods. In the first method, bacterial cells containing individual clones are robotically picked and arranged on a nylon membrane. The membrane is placed on LB agar containing selective agent (carbenicillin, kanamycin, ampicillin, or chloramphenicol depending on the vector used) and incubated at 37C for 16 hr. The membrane is removed from the agar and consecutively placed colony side up in 10% SDS, denaturing solution (1.5 M NaCl, 0.5 M NaOH), neutralizing solution (1.5 M NaCl, 1 M Tris, pH 8.0), and twice in 2xSSC for 10 min each. The membrane is then UV irradiated in a STRATALINKER UV-crosslinker (Stratagene).

In the second method, polynucleotides are amplified from bacterial vectors by thirty cycles of PCR using primers complementary to vector sequences flanking the insert. PCR amplification increases a starting concentration of 1-2 ng nucleic acid to a final quantity greater than 5 µg. Amplified nucleic acids from about 400 bp to about 5000 bp in length are purified using SEPHACRYL-400 beads (APB). Purified nucleic acids are arranged on a nylon membrane manually or using a dot/slot blotting manifold and suction device and are immobilized by denaturation, neutralization, and UV irradiation as described above. Purified nucleic acids are robotically arranged and immobilized on polymer-coated glass slides using the procedure described in USPN 5,807,522. Polymer-coated slides are prepared by cleaning glass microscope slides (Corning Life Sciences) by ultrasound in 0.1% SDS and acetone, etching in 4% hydrofluoric acid (VWR Scientific Products, West Chester PA), coating with 0.05% aminopropyl silane (Sigma Aldrich) in 95% ethanol, and curing in a 110C oven. The slides are washed extensively with distilled water between and after treatments. The nucleic acids are arranged on the slide and then immobilized by exposing the array to UV irradiation using a STRATALINKER UV-crosslinker (Stratagene). Arrays are then washed at room temperature in 0.2% SDS and rinsed three times in distilled water. Non-specific binding sites are blocked by incubation of arrays in 0.2% casein in phosphate buffered saline (PBS; Tropix, Bedford MA) for 30 min at 60C; then the arrays are washed in 0.2% SDS and rinsed in distilled water as before.

### 30 Probe Preparation for Membrane Hybridization

Hybridization probes derived from the polynucleotides of the Sequence Listing are employed for screening cDNAs, mRNAs, or genomic DNA in membrane-based hybridizations. Probes are prepared by diluting the polynucleotides to a concentration of 40-50 ng in 45 µl TE buffer, denaturing by heating to 100C for five min, and briefly centrifuging. The denatured polynucleotide is then added to a REDIPRIME

tube (APB), gently mixed until blue color is evenly distributed, and briefly centrifuged. Five  $\mu$ l of [ $^{32}$ P]dCTP is added to the tube, and the contents are incubated at 37C for 10 min. The labeling reaction is stopped by adding 5  $\mu$ l of 0.2M EDTA, and probe is purified from unincorporated nucleotides using a PROBEQUANT G-50 microcolumn (APB). The purified probe is heated to 100C for five min, snap  
 5 cooled for two min on ice, and used in membrane-based hybridizations as described below.

#### Probe Preparation for QPCR

Probes for the QPCR were prepared according to the ABI protocol.

#### Probe Preparation for Polymer Coated Slide Hybridization

The following method was used for the preparation of probes for the microarray analyses  
 10 presented in Tables 1 and 2 Hybridization probes derived from mRNA isolated from samples are employed for screening polynucleotides of the Sequence Listing in array-based hybridizations. Probe is prepared using the GEMbright kit (Incyte Genomics) by diluting mRNA to a concentration of 200 ng in 9  $\mu$ l TE buffer and adding 5  $\mu$ l 5x buffer, 1  $\mu$ l 0.1 M DTT, 3  $\mu$ l Cy3 or Cy5 labeling mix, 1  $\mu$ l RNase inhibitor, 1  $\mu$ l reverse transcriptase, and 5  $\mu$ l 1x yeast control mRNAs. Yeast control mRNAs are  
 15 synthesized by in vitro transcription from noncoding yeast genomic DNA (W. Lei, unpublished). As quantitative controls, one set of control mRNAs at 0.002 ng, 0.02 ng, 0.2 ng, and 2 ng are diluted into reverse transcription reaction mixture at ratios of 1:100,000, 1:10,000, 1:1000, and 1:100 (w/w) to sample mRNA respectively. To examine mRNA differential expression patterns, a second set of control mRNAs are diluted into reverse transcription reaction mixture at ratios of 1:3, 3:1, 1:10, 10:1, 1:25, and 25:1  
 20 (w/w). The reaction mixture is mixed and incubated at 37C for two hr. The reaction mixture is then incubated for 20 min at 85C, and probes are purified using two successive CHROMA SPIN+TE 30 columns (Clontech, Palo Alto CA). Purified probe is ethanol precipitated by diluting probe to 90  $\mu$ l in DEPC-treated water, adding 2  $\mu$ l 1mg/ml glycogen, 60  $\mu$ l 5 M sodium acetate, and 300  $\mu$ l 100% ethanol. The probe is centrifuged for 20 min at 20,800xg, and the pellet is resuspended in 12  $\mu$ l resuspension  
 25 buffer, heated to 65C for five min, and mixed thoroughly. The probe is heated and mixed as before and then stored on ice. Probe is used in high density array-based hybridizations as described below.

#### In situ Hybridization

The following method was used in the analyses performed in Figures 7 and 8. In situ  
 hybridization was used to determine the expression of HG38 in sectioned tissue. With the digoxigenin  
 30 protocol, fresh cryosections, 10 microns thick, were removed from the freezer, immediately immersed in 4% paraformaldehyde for 10 min, rinsed in PBS, and acetylated in 0.1 M TEA, pH 8.0, containing 0.25% (v/v) acetic anhydride. After the tissue equilibrated in 5 x SSC, it was prehybridized in hybridization buffer (50% formamide, 5 x SSC, 1 x Denhardt's solution, 10% dextran sulfate, and 1 mg/ml herring sperm DNA).

Digoxigenin-labeled HG38-specific RNA probes, sense and antisense nucleotides selected from the polynucleotide of SEQ ID NO:1 were produced using PCR. Approximately 500 ng/ml of probe was used in overnight hybridizations at 65C in hybridization buffer. Following hybridization, the sections were rinsed for 30 min in 2 x SSC at room temperature, 1 hr in 2 x SSC at 65C, and 1 hr in 0.1 x SSC at 5 65C. The sections were equilibrated in PBS, blocked for 30 min in 10% DIG kit blocker (Roche Molecular Biochemicals, Indianapolis IN) in PBS, then incubated overnight at 4C in 1:500 anti-DIG-AP. The following day, the sections were rinsed in PBS, equilibrated in detection buffer (0.1 M Tris, 0.1 M NaCl, 50 mM MgCl<sub>2</sub>, pH 9.5), and then incubated in detection buffer containing 0.175 mg/ml NBT and 0.35 mg/ml BCIP. The reaction was terminated in TB, pH 8. Tissue sections were counterstained with 1 10 µg/ml DAPI and mounted in VECTASHIELD (Vector Laboratory, Burlingame CA).

#### Membrane-based Hybridization

Membranes are pre-hybridized in hybridization solution containing 1% Sarkosyl and 1x high phosphate buffer (0.5 M NaCl, 0.1 M Na<sub>2</sub>HPO<sub>4</sub>, 5 mM EDTA, pH 7) at 55C for two hr. The probe, diluted in 15 ml fresh hybridization solution, is then added to the membrane. The membrane is hybridized 15 with the probe at 55C for 16 hr. Following hybridization, the membrane is washed for 15 min at 25C in 1mM Tris (pH 8.0), 1% Sarkosyl, and four times for 15 min each at 25C in 1mM Tris (pH 8.0). To detect hybridization complexes, XOMAT-AR film (Eastman Kodak, Rochester NY) is exposed to the membrane overnight at -70C, developed, and examined visually.

#### Polymer Coated Slide-based Hybridization

20 The following method was used in the microarray analyses presented in Tables 1 and 2. Probe is heated to 65C for five min, centrifuged five min at 9400 rpm in a 5415C microcentrifuge (Eppendorf Scientific, Westbury NY), and then 18 µl is aliquoted onto the array surface and covered with a coverslip. The arrays are transferred to a waterproof chamber having a cavity just slightly larger than a microscope slide. The chamber is kept at 100% humidity internally by the addition of 140 µl of 5xSSC in a corner of 25 the chamber. The chamber containing the arrays is incubated for about 6.5 hr at 60C. The arrays are washed for 10 min at 45C in 1xSSC, 0.1% SDS, and three times for 10 min each at 45C in 0.1xSSC, and dried.

Hybridization reactions are performed in absolute or differential hybridization formats. In the absolute hybridization format, probe from one sample is hybridized to array elements, and signals are 30 detected after hybridization complexes form. Signal strength correlates with probe mRNA levels in the sample. In the differential hybridization format, differential expression of a set of genes in two biological samples is analyzed. Probes from the two samples are prepared and labeled with different labeling moieties. A mixture of the two labeled probes is hybridized to the array elements, and signals are examined under conditions in which the emissions from the two different labels are individually

detectable. Elements on the array that are hybridized to equal numbers of probes derived from both biological samples give a distinct combined fluorescence (Shalon WO95/35505).

Hybridization complexes are detected with a microscope equipped with an Innova 70 mixed gas 10 W laser (Coherent, Santa Clara CA) capable of generating spectral lines at 488 nm for excitation of Cy3 and at 632 nm for excitation of Cy5. The excitation laser light is focused on the array using a 20X microscope objective (Nikon, Melville NY). The slide containing the array is placed on a computer-controlled X-Y stage on the microscope and raster-scanned past the objective with a resolution of 20 micrometers. In the differential hybridization format, the two fluorophores are sequentially excited by the laser. Emitted light is split, based on wavelength, into two photomultiplier tube detectors (PMT R1477, 10 Hamamatsu Photonics Systems, Bridgewater NJ) corresponding to the two fluorophores. Filters positioned between the array and the photomultiplier tubes are used to separate the signals. The emission maxima of the fluorophores used are 565 nm for Cy3 and 650 nm for Cy5. The sensitivity of the scans is calibrated using the signal intensity generated by the yeast control mRNAs added to the probe mix. A specific location on the array contains a complementary DNA sequence, allowing the intensity of the 15 signal at that location to be correlated with a weight ratio of hybridizing species of 1:100,000.

The output of the photomultiplier tube is digitized using a 12-bit RTI-835H analog-to-digital (A/D) conversion board (Analog Devices, Norwood MA) installed in an IBM-compatible PC computer. The digitized data are displayed as an image where the signal intensity is mapped using a linear 20-color transformation to a pseudocolor scale ranging from blue (low signal) to red (high signal). The data is also 20 analyzed quantitatively. Where two different fluorophores are excited and measured simultaneously, the data are first corrected for optical crosstalk (due to overlapping emission spectra) between the fluorophores using the emission spectrum for each fluorophore. A grid is superimposed over the fluorescence signal image such that the signal from each spot is centered in each element of the grid. The fluorescence signal within each element is then integrated to obtain a numerical value corresponding to the 25 average intensity of the signal. The software used for signal analysis is the GEMTOOLS program (Incyte Genomics).

#### QPCR Analysis

For QPCR, cDNA was synthesized from 1 ug total RNA in a 25 ul reaction with 100 units M-MLV reverse transcriptase (Ambion, Austin TX), 0.5 mM dNTPs (Epicentre, Madison WI), and 40 ng/ml 30 random hexamers (Fisher Scientific, Chicago IL). Reactions were incubated at 25C for 10 minutes, 42C for 50 minutes, and 70C for 15 minutes, diluted to 500 ul, and stored at -30C. Alternatively, normal tissues were purchased from Clontech (Palo Alto CA) and Clinomics. PCR primers and probes (5' 6-FAM-labeled, 3' TAMRA) were designed using PRIMER EXPRESS 1.5 software (ABI) and synthesized by Biosearch Technologies (Novato CA) or ABI.

QPCR reactions were performed using an PRISM 7700 detection system (ABI) in 25 ul total volume with 5 ul cDNA template, 1x TAQMAN UNIVERSAL PCR master mix (ABI), 100 nM each PCR primer, 200 nM probe, and 1x VIC-labeled beta-2-microglobulin endogenous control (ABI). Reactions were incubated at 50C for 2 minutes, 95C for 10 minutes, followed by 40 cycles of incubation at 95C for 15 seconds and 60C for 1 minute. Emissions were measured once every cycle, and results were analyzed using SEQUENCE DETECTOR 1.7 software (ABI) and fold differences, relative concentration of mRNA as compared to standards, were calculated using the comparative  $C_T$  method (ABI User Bulletin #2). QPCR was used to produce the data for Figures 2-6.

### VIII Complementary Molecules

Antisense molecules complementary to the cDNA, from about 5 bp to about 5000 bp in length, are used to detect or inhibit gene expression. Detection is described in Example VII. To inhibit transcription by preventing promoter binding, the complementary molecule is designed to bind to the most unique 5' sequence and includes nucleotides of the 5' UTR upstream of the initiation codon of the open reading frame. Complementary molecules include genomic sequences (such as enhancers or introns) and are used in triple helix base pairing to compromise the ability of the double helix to open sufficiently for the binding of polymerases, transcription factors, or regulatory molecules. To inhibit translation, a complementary molecule is designed to prevent ribosomal binding to the mRNA encoding the protein.

Complementary molecules are placed in expression vectors and used to transform a cell line to test efficacy; into an organ, tumor, synovial cavity, or the vascular system for transient or short term therapy; or into a stem cell, zygote, or other reproducing lineage for long term or stable gene therapy. Transient expression lasts for a month or more with a non-replicating vector and for three months or more if elements for inducing vector replication are used in the transformation/expression system.

Stable transformation of dividing cells with a vector encoding the complementary molecule produces a transgenic cell line, tissue, or organism (USPN 4,736,866). Those cells that assimilate and replicate sufficient quantities of the vector to allow stable integration also produce enough complementary molecules to compromise or entirely eliminate activity of the polynucleotide encoding the protein.

### IX Production of Specific Antibodies

Purification using polyacrylamide gel electrophoresis or similar techniques is used to isolate protein for immunization of hosts or host cells to produce antibodies using standard protocols.

Alternatively, the amino acid sequence of the protein is analyzed using LASERGENE software (DNASTAR) to determine regions of high immunogenicity. A peptide with high immunogenicity is cleaved, recombinantly-produced, or synthesized and used to raise antibodies by means known to those of skill in the art. Methods for selection of appropriate antigenic determinants such as those near the



C-terminus or in hydrophilic regions are well described in the art (Ausubel, supra, Chap. 11).

Oligopeptides of about 15 residues in length are synthesized using an 431A peptide synthesizer (ABI) using FMOC chemistry and coupled to carriers such as BSA, thyroglobulin, or KLH (Sigma-Aldrich) by reaction with N-maleimidobenzoyl-N-hydroxysuccinimide ester to increase immunogenicity. The coupled  
5 peptide is then used to immunize the host. Rabbits are immunized with the oligopeptide-KLH complex in complete Freund's adjuvant. Resulting antisera are tested for anti-peptide activity by binding the peptide to a substrate, blocking with 1% BSA, reacting with rabbit antisera, washing, and reacting with radio-iodinated goat anti-rabbit IgG.

#### **X Immunopurification Using Antibodies**

10 Naturally occurring or recombinantly produced protein is purified by immunoaffinity chromatography using antibodies which specifically bind the protein. An immunoaffinity column is constructed by covalently coupling the antibody to CNBr-activated SEPHAROSE resin (APB). Media containing the protein is passed over the immunoaffinity column, and the column is washed using high ionic strength buffers in the presence of detergent to allow preferential absorbance of the protein. After  
15 coupling, the protein is eluted from the column using a buffer of pH 2-3 or a high concentration of urea or thiocyanate ion to disrupt antibody/protein binding, and the purified protein is collected.

#### **XI Western Analysis**

##### Electrophoresis and Blotting

Samples containing protein are mixed in 2 x loading buffer, heated to 95 C for 3-5 min, and loaded  
20 on 4-12% NUPAGE Bis-Tris precast gel (Invitrogen). The gel is electrophoresed in 1 x MES or MOPS running buffer (Invitrogen) at 200 V for approximately 45 min on a (apparatus, supplier) until the RAINBOW marker (APB) has resolved, and dye front approaches the bottom of the gel. The gel and its supports are removed from the apparatus and soaked in 1 x transfer buffer (Invitrogen) with 10% methanol for a few minutes; and the PVDF membrane soaked in 100% methanol for a few seconds to activate it.  
25 The membrane, the gel, and supports are placed on the transfer apparatus (machine, supplier) and a constant current of 350 mAmps is applied for 90 min.

##### Conjugation with Antibody and Visualization

After the proteins are transferred to the membrane, it is blocked in 5% (w/v) non-fat dry milk in 1 x phosphate buffered saline (PBS) with 0.1% Tween 20 detergent (blocking buffer) on a rotary shaker  
30 (supplier) for at least 1 hr at room temperature or at 4° overnight. After blocking, the buffer is removed, and 10 ml of primary antibody in blocking buffer is added and incubated on the rotary shaker for 1 hr at room temperature or overnight at 4 C. The membrane is washed 3 x for 10 min each with PBS-Tween (PBST), and secondary antibody, conjugated to horseradish peroxidase, is added at a 1:3000 dilution in 10

ml blocking buffer. The membrane and solution are shaken for 30 min at room temperature and then washed 3 x for 10 min each with PBST.

The wash solution is carefully removed, and the membrane moistened with ECL+ chemiluminescent detection system (APB) and incubated for approximately 5 min. The membrane is placed, protein side down, on plastic film (product, supplier) and developed for approximately 30 seconds.

## **XII Antibody Arrays**

### **Protein:protein interactions**

In an alternative to yeast two hybrid system analysis of proteins, an antibody array can be used to study protein-protein interactions and phosphorylation. A variety of protein ligands are immobilized on a membrane using methods well known in the art. The array is incubated in the presence of cell lysate until protein:antibody complexes are formed. Proteins of interest are identified by exposing the membrane to an antibody specific to the protein of interest. In the alternative, a protein of interest is labeled with digoxigenin (DIG) and exposed to the membrane; then the membrane is exposed to anti-DIG antibody which reveals where the protein of interest forms a complex. The identity of the proteins with which the protein of interest interacts is determined by the position of the protein of interest on the membrane.

### **Proteomic Profiles**

Antibody arrays can also be used for high-throughput screening of recombinant antibodies. Bacteria containing antibody genes are robotically-picked and gridded at high density (up to 18,342 different double-spotted clones) on a filter. Up to 15 antigens at a time are used to screen for clones to identify those that express binding antibody fragments. These antibody arrays can also be used to identify proteins which are differentially expressed in samples (de Wildt, *supra*)

## **XIII Screening Molecules for Specific Binding with the polynucleotide or Protein**

The polynucleotide, or fragments thereof, or the protein, or portions thereof, are labeled with <sup>32</sup>P-dCTP, Cy3-dCTP, or Cy5-dCTP (APB), or with BIODIPY or FITC (Molecular Probes), respectively. Libraries of candidate molecules or compounds previously arranged on a substrate are incubated in the presence of labeled polynucleotide or protein. After incubation under conditions for either a nucleic acid or amino acid sequence, the substrate is washed, and any position on the substrate retaining label, which indicates specific binding or complex formation, is assayed, and the ligand is identified. Data obtained using different concentrations of the nucleic acid or protein are used to calculate affinity between the labeled nucleic acid or protein and the bound molecule.

**XIV Two-Hybrid Screen**

A yeast two-hybrid system, MATCHMAKER LexA Two-Hybrid system (Clontech Laboratories), is used to screen for peptides that bind the protein of the invention. A polynucleotide encoding the protein is inserted into the multiple cloning site of a pLexA vector, ligated, and transformed into *E. coli*. cDNA, prepared from mRNA, is inserted into the multiple cloning site of a pB42AD vector, ligated, and transformed into *E. coli* to construct a cDNA library. The pLexA plasmid and pB42AD-cDNA library constructs are isolated from *E. coli* and used in a 2:1 ratio to co-transform competent yeast EGY48[p8op-lacZ] cells using a polyethylene glycol/lithium acetate protocol. Transformed yeast cells are plated on synthetic dropout (SD) media lacking histidine (-His), tryptophan (-Trp), and uracil (-Ura), and incubated at 30°C until the colonies have grown up and are counted. The colonies are pooled in a minimal volume of 1x TE (pH 7.5), replated on SD/-His/-Leu/-Trp/-Ura media supplemented with 2% galactose (Gal), 1% raffinose (Raf), and 80 mg/ml 5-bromo-4-chloro-3-indolyl  $\beta$ -d-galactopyranoside (X-Gal), and subsequently examined for growth of blue colonies. Interaction between expressed protein and polynucleotide fusion proteins activates expression of a LEU2 reporter gene in EGY48 and produces colony growth on media lacking leucine (-Leu). Interaction also activates expression of  $\beta$ -galactosidase from the p8op-lacZ reporter construct that produces blue color in colonies grown on X-Gal.

Positive interactions between expressed protein and polynucleotide fusion proteins are verified by isolating individual positive colonies and growing them in SD/-Trp/-Ura liquid medium for 1 to 2 days at 30°C. A sample of the culture is plated on SD/-Trp/-Ura media and incubated at 30°C until colonies appear. The sample is replica-plated on SD/-Trp/-Ura and SD/-His/-Trp/-Ura plates. Colonies that grow on SD containing histidine but not on media lacking histidine have lost the pLexA plasmid. Histidine-requiring colonies are grown on SD/Gal/Raf/X-Gal/-Trp/-Ura, and white colonies are isolated and propagated. The pB42AD-cDNA plasmid, which contains a cDNA encoding a protein that physically interacts with the protein, is isolated from the yeast cells and characterized.

**25 XV GPCR Activity Assay**

The GPCR encoded by SEQ ID NO:2 may be expressed in heterologous expression systems and its biological activity tested utilizing the purinergic receptor system ( $P_{2U}$ ) as published by Erb *et al.* (1993, Proc Natl Acad Sci 90:10449-53). Because cultured K562 human leukemia cells lack  $P_{2U}$  receptors, they can be transfected with expression vectors containing either normal or chimeric  $P_{2U}$  and loaded with fura- $\alpha$ , fluorescent probe for  $Ca^{++}$ . Activation of properly assembled and functional extracellular SP-transmembrane/intracellular  $P_{2U}$  receptors with extracellular UTP or ATP mobilizes intracellular  $Ca^{++}$  which reacts with fura- $\alpha$  and is measured spectrofluorometrically. Bathing the transfected K562 cells in microwells containing appropriate ligands will trigger binding and fluorescent activity defining effectors

of SP. Once ligand and function are established, the  $P_{2U}$  system is useful for defining antagonists or inhibitors which block binding and prevent such fluorescent reactions.

## **XVI Cell Transformation Assays**

### Colony-formation Assay in Soft Agar

5 The ability of transformed cells to grow in an anchorage-independent manner is measured by the ability of the cells to form colonies in soft agar (0.35%). The assay is conducted in 12-well culture plates where each well is coated with a solid 0.7% Noble agar (Fisher Scientific, Atlanta GA) in cell growth media. A 3.5% agar solution in PBS is prepared, autoclaved, microwaved and kept liquid in a 55 C water bath with shaking. The agar is diluted 1:5 to 0.7% with an appropriate cell growth media, and 0.5 ml of  
10 the diluted agar added to each well of the plate. Culture plates are kept at room temperature for about 15 minutes or until the agar solidifies.

Trypsinized cells are diluted to 200 to 4000 cells/ml in growth medium and 0.25 ml of diluted cells is mixed with 2 ml warm 0.35% agar. The diluted cells are added to a well of the culture plate; duplicate wells are prepared for each cell concentration. The plates are allowed to cool for about 30 min  
15 at room temperature and then transferred to an incubator at 37 C. After a 1-2 week incubation period, colonies are counted under an inverted, phase contrast microscope. Colony forming efficiency is determined as the percentage colonies formed/total number of cells plated.

### Apoptosis/Survival Assay

The ability of transformed cells to evade apoptosis (programmed cell death) and survive may be  
20 measured in an assay in which apoptosis or survival of cultured cells is determined by FACS analysis using a double-staining method with Annexin V and propidium iodide (PI). Annexin V serves as a marker for apoptotic cells by binding to phosphatidyl serine, a cell surface marker for apoptosis. Counterstaining with PI allows differentiation between apoptotic cells, which are Annexin V positive and PI negative, and necrotic cells, which are Annexin V and PI positive. Apoptosis is measured between 0-24 hrs of culture,  
25 and cell survival is measured between 24-96 hrs of culture.

Alternatively, the direct effect of a secreted protein, such as HG3, on apoptosis/cell survival may be measured in cultured human vascular endothelial cells (HMVEC) following treatment of HMVEC cells with HG38, or infection of the cells with a recombinant adenovirus containing the cDNA encoding HG38. Apoptosis/survival of the HMVEC cells is measured as described above.

### 30 Tissue Invasion and Metastasis Assay

Cell migration and tissue invasion by transformed tumor cells is determined using the BICOAT Angiogenesis system (BD Biosciences, Franklin Lakes NJ) as described by the manufacturer. The assay is carried out in a BD FALCON multiwell insert plate containing an 8  $\mu$ m pore size BD FLUOROBLOK polyethylene terephthalate membrane uniformly coated with a reconstituted BD MATRIGEL basement

membrane matrix and inserted into a non-treated multiwell receiver plate. The system provides a barrier to passive diffusion of cells through the membrane but allows active migration by invasive tumor cells. After cells in appropriate culture medium are incubated in the upper portion of the chamber for a suitable period of time, any cells appearing on the underside of the membrane are quantitated. Since the membrane  
5 blocks the transmission of light from 490 to 700nm, cells traversing the membrane are detected by their fluorescence which is proportionate to cell number.

All patents and publications mentioned in the specification are incorporated by reference herein. Various modifications and variations of the described method and system of the invention will be apparent  
10 to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the described modes for carrying out the invention that are obvious to those skilled in the field of molecular biology or related fields are intended to be within the scope of the following claims.

15

Table 1

Normal		Tumor	
Cy5/Cy3	Dn ID	Description	Dn ID
2.50	Pool	Colon Tumor	Dn3757
2.20	Pool	Colon Tumor	Dn3756
2.00	Pool	Colon Tumor,	Dn3756
2.00	Dn3583	ColonTumor, Adenoma	Dn3583
2.70	Dn3583	ColonTumor, Adenoma	Dn3583
2.00	Dn3647	Colon Tumor	Dn3647
2.60	Dn3647	Colon Tumor	Dn3647
1.80	Dn3579	Colon Tumor, AdenoCA	Dn3579
1.70	Dn3582	Colon Tumor, AdenoCa	Dn3582
2.80	Dn3839	Colon Tumor, AdenoCA	Dn3839
3.90	Dn9573	Colon Tumor, AdenoCA	Dn9573
4.00	Dn9573	Colon Tumor, AdenoCA,	Dn9573
1.60	Dn9574	Colon Tumor, Rectum, AdenoCA	Dn9574
1.20	Dn9574	Colon Tumor, Rectum, AdenoCA	Dn9574
1.60	Dn9576	Colon Tumor, Rectum, AdenoCA	Dn9576
2.10	Dn9576	Colon Tumor, Rectum, AdenoCA	Dn9576
1.40	Dn9577	Colon Tumor, Rectum, AdenoCA	Dn9577
1.40	Dn8403	Colon, Tumor, Ulcerative AdenoCA	Dn8403
1.20	Dn8403	Colon, Tumor, Ulcerative AdenoCA	Dn8403

TABLE 2

Cy5/C Normal		Tumor Description		umor Dn ID	
y3	Dn ID				
-1.40	Dn7168	Human, Lung Tumor,	Squamous Cell CA,		Dn7168
-1.20	Dn7168	Human, Lung Tumor,	Squamous Cell CA,		Dn7168
-1.20	Dn7164	Human, Lung Tumor,	Carcinoid,		Dn7164
-1.40	Dn7164	Human, Lung Tumor,	Carcinoid,		Dn7164
-1.50	Dn7162	Human, Lung Tumor,	Large Cell Endocrine C		Dn7162
-2.00	Dn7162	Human, Lung Tumor,	Large Cell Endocrine C		Dn7162
1.80	Dn7178	Human, Lung Tumor,	Left Upper Lobe, Squam		Dn7178
-1.60	Dn7176	Human, Lung Tumor,	Right Middle Lower Lob		Dn7176
2.20	Dn7173	Human, Lung Tumor,	Right, Squamous Cell C		Dn7173
3.80	Dn7173	Human, Lung Tumor,	Right, Squamous Cell C		Dn7173
1.30	Dn7190	Human, Lung Tumor,	Left, Squamous Cell CA		Dn7190
1.80	Pool	Human, Lung Tumor,	Right Middle Lobe, Aty		Dn7186
2.80	Dn7963	Human, Lung Tumor,	Non-Small Cell Lung CA		Dn7963
-1.40	Dn7191	Human, Lung Tumor,	Left, Squamous Cell CA		Dn7191
1.60	Dn5800	Human, Lung Tumor,	Squamous Cell CA,		Dn5800
1.60	Dn5800	Human, Lung Tumor,	Squamous Cell CA,		Dn5800
1.10	Dn5799	Human, Lung Tumor,	AdenOCA,		Dn5799
1.00	Dn5799	Human, Lung Tumor,	AdenOCA,		Dn5799
-1.20	Dn5798	Human, Lung Tumor,	AdenOCA,		Dn5798
-1.10	Dn5798	Human, Lung Tumor,	AdenOCA,		Dn5798
2.10	Dn5797	Human, Lung Tumor,	Squamous Cell CA,		Dn5797
2.40	Dn5797	Human, Lung Tumor,	Squamous Cell CA,		Dn5797
1.80	Dn5796	Human, Lung Tumor,	Squamous Cell CA,		Dn5796
-1.20	Dn5795	Human, Lung Tumor,	AdenOCA,		Dn5795
1.30	Dn5793	Human, Lung Tumor,	Squamous Cell CA,		Dn5793
1.20	Dn5793	Human, Lung Tumor,	Squamous Cell CA,		Dn5793

What is claimed is:

1. A method for using a polynucleotide to detect a colon or lung cancer comprising:
  - a) hybridizing a composition comprising the polynucleotide of SEQ ID NO:2, or the complement thereof, and a labeling moiety, to nucleic acids of a sample of colon or lung tissue under conditions to  
5 form at least one hybridization complex;
  - b) detecting hybridization complex formation; and
  - c) comparing complex formation to a standard, wherein the comparison reflects differential expression of the polynucleotide in the sample relative to the standard and is diagnostic of a colon or lung cancer.
- 10 2. The method of claim 1 further comprising amplifying the nucleic acids of the sample prior to hybridization.
3. The method of claim 1 wherein the composition is attached to a substrate.
4. A method for detecting a colon or lung cancer, the method comprising:
  - a) performing an assay to determine the amount of the protein of SEQ ID NO:1 in a sample of  
15 colon or lung tissue; and
  - b) comparing the amount of protein to a standard, thereby detecting expression of the protein in the sample, wherein differential expression of the protein in the sample when compared with the standard is diagnostic of a colon or lung cancer.
5. The method of claim 4 wherein the assay is selected from antibody arrays, enzyme-linked  
20 immunosorbent assays, fluorescence-activated cell sorting, two dimensional-polyacrylamide gel electrophoresis and scintillation counting, radioimmunoassays, and western analysis.
6. The method of claim 4 comprising:
  - a) combining an antibody specific for SEQ ID NO:1 with a sample under conditions which allow the formation of antibody:protein complexes;
  - 25 b) detecting complex formation wherein complex formation indicates expression of the protein in the sample; and
  - c) comparing complex formation with a standard, wherein differential expression of the protein between the sample and the standard is diagnostic of a colon or lung cancer.
7. A method for treating a colon or lung cancer comprising administering to a subject in need of  
30 therapeutic intervention the antibody of claim 6.
9. A method for delivering a therapeutic agent to a colon cancer cell comprising:
  - a) attaching the therapeutic agent to the antibody of claim 6; and



b) administering the antibody to a subject in need of therapeutic intervention, wherein the antibody specifically binds the protein having the amino acid sequence of SEQ ID NO:1 thereby delivering the therapeutic agent to the cell.

10. A method for treating a colon or lung cancer comprising administering to a subject in need of  
5 therapeutic intervention an antagonist of the protein of claim 4.

	11	20	29	38	47	56
5'CC	CCT ACT TCG GGC ACC ATG GAC ACC TCC CGG CTC GGT GTG CTC CTG TCC TTG CCT					
	---	---	---	---	---	---
	Met Asp Thr Ser Arg Leu Gly Val Leu Ser Leu Pro					
	65	74	83	92	101	110
GTG CTG CTG CAG CTG GCG ACC GGG GGC AGC TCT CCC AGG TCT GGT GTG TTG CTG						
---	---	---	---	---	---	---
Val Leu Leu Gln Leu Ala Thr Gly Gly Ser Ser Pro Arg Ser Gly Val Leu Leu						
119	128	137	146	155		164
AGG GGC TGC CCC ACA CAC CAC TGT CAT TGC GAG CCC GAC GGC AGG ATG TTG CTC AGG						
---	---	---	---	---	---	---
Arg Gly Cys Pro Thr His Cys His Cys Glu Pro Asp Gly Arg Met Leu Leu Arg						
173	182	191	200	209		218
GTG GAC TGC TCC GAC CTG GGG CTC TCG GAG GAG CTG CCT TCC AAC CTC AGC GTC TTC						
---	---	---	---	---	---	---
Val Asp Cys Ser Asp Leu Gly Leu Ser Glu Leu Pro Ser Asn Leu Ser Val Phe						
227	236	245	254	263		272
ACC TCC TAC CTA GAC CTC AGT ATG AAC AAC AATC ATC AGT CAG CTG CTC CCG AAT CCC						
---	---	---	---	---	---	---
Thr Ser Tyr Leu Asp Leu Ser Met Asn Asn Ile Ser Gln Leu Leu Pro Asn Pro						
281	290	299	308	317		326
CTG CCC AGT CTC CGC TTC CTG GAG GAG TTA CGT CTT GCG GGA AAC GCT CTG ACA						
---	---	---	---	---	---	---
Leu Pro Ser Leu Arg Phe Leu Glu Glu Leu Arg Leu Ala Gly Asn Ala Leu Thr						

2/17

335	344	353	362	371	380
TAC ATT CCC AAG GGA GCA TTC ACT GGC CTT TAC AGT CTT AAA GTT CTT ATG CTG					
---	---	---	---	---	---
Tyr Ile Pro Lys Gly Ala Phe Thr Gly Leu Tyr Ser Leu Lys Val Leu Met Leu					
389	398	407	416	425	434
CAG AAT AAT CAG CTA AGA CAC GTA CCC ACA GAA GCT CTG CAG AAT TTG CGA AGC					
---	---	---	---	---	---
Gln Asn Asn Gln Leu Arg His Val Pro Thr Glu Ala Leu Gln Asn Leu Arg Ser					
443	452	461	470	479	488
CTT CAA TCC CTG CTG CGT CTG GAT GAT GCT AAC CAC CAC ATC AGC TAT GTG CCC CCA AGC TGT					
---	---	---	---	---	---
Leu Gln Ser Leu Arg Leu Asp Ala Asn His Ile Ser Tyr Val Pro Pro Ser Cys					
497	506	515	524	533	542
TTC AGT GGC CTG CAT TCC CTG CTG AGG CAC CAC CTG TGG CTG GAT GAC AAT GCG TTA ACA					
---	---	---	---	---	---
Phe Ser Gly Leu His Ser Leu Arg His Leu Tyr Trp Leu Asp Asp Asn Ala Leu Thr					
551	560	569	578	587	596
GAA ATC CCC GTC CAG GCT TTT AGA AGT TTA TCG GCA TTG CAA GCC ATG ACC TTG					
---	---	---	---	---	---
Glu Ile Pro Val Gln Ala Phe Arg Ser Leu Ser Ala Leu Gln Ala Met Thr Leu					
605	614	623	632	641	650
GCC CTG AAC AAA ATA CAC CAC ATA CCA GAC TAT GCC TTT GGA AAC CTC TCC AGC					
---	---	---	---	---	---
Ala Leu Asn Lys Ile His His Ile Pro Asp Tyr Ala Phe Gly Asn Leu Ser Ser					

FIGURE 1B

3/17

659	668	677	686	695	704
TTG GTA GTT CTA CAT CTC CAT AAC AAT AGA ATC CAC TCC CTG GGA AAG AAA TGC					
---	---	---	---	---	---
Leu Val Val Leu His Leu His Asn Asn Arg Ile His Ser Leu Gly Lys Lys Cys					
713	722	731	740	749	758
TTT GAT GGG CTC CAC AGC CTA GAG ACT TTA GAT TTA AAT TAC AAT AAC CTT GAT					
---	---	---	---	---	---
Phe Asp Gly Leu His Ser Leu His Ser Leu Asp Leu Asn Tyr Asn Asn Leu Asp					
767	776	785	794	803	812
GAA TTC CCC ACT GCA ATT AGG ACA CTC TCC AAC CTT AAA GAA CTA GGA TTT CAT					
---	---	---	---	---	---
Glu Phe Pro Thr Ala Ile Arg Thr Leu Ser Asn Leu Lys Glu Leu Gly Phe His					
821	830	839	848	857	866
AGC AAC AAT ATC AGG TCG ATA CCT GAG AAA GCA TTT GTA GGC AAC CCT TCT CTT					
---	---	---	---	---	---
Ser Asn Asn Ile Arg Ser Ile Pro Glu Lys Ala Phe Val Gly Asn Pro Ser Leu					
875	884	893	902	911	920
ATT ACA ATA CAT TTC TAT GAC AAT CCC ATC CAA TTT GTT GGG AGA TCT GCT TTT					
---	---	---	---	---	---
Ile Thr Ile His Phe Tyr Asp Asn Pro Ile Gln Phe Val Gly Arg Ser Ala Phe					
929	938	947	956	965	974
CAA CAT TTA CCT GAA CTA AGA ACA CTG ACT CTG AAT GGT GCC TCA CAA ATA ACT					
---	---	---	---	---	---
Gln His Leu Pro Glu Leu Arg Thr Leu Thr Leu Asn Gly Ala Ser Gln Ile Thr					

FIGURE 1C

4/17

983	992	1001	1010	1019	1028
GAA TTT CCT GAT TTA ACT GGA ACT GCA AAC CTG GAG AGT CTG ACT TTA ACT GGA					
---	---	---	---	---	---
Glu Phe Pro Asp Leu Thr Gly Thr Ala Asn Leu Glu Ser Leu Thr Leu Thr Gly					
1037	1046	1055	1064	1073	1082
GCA CAG ATC TCA TCT CTT CCT CAA ACC GTC TGC AAT CAG TTA CCT AAT CTC CAA					
---	---	---	---	---	---
Ala Gln Ile Ser Ser Leu Pro Gln Thr Val Cys Asn Gln Leu Pro Asn Leu Gln					
1091	1100	1109	1118	1127	1136
GTG CTA GAT CTG TCT TAC AAC CTA TTA GAA GAT TTA CCC AGT TTT TCA GTC TGC					
---	---	---	---	---	---
Val Leu Asp Leu Ser Tyr Asn Leu Leu Glu Asp Leu Pro Ser Phe Ser Val Cys					
1145	1154	1163	1172	1181	1190
CAA AAG CTT CAG AAA ATT GAC CTA AGA CAT AAT GAA ATC TAC GAA ATT AAA GTT					
---	---	---	---	---	---
Gln Lys Leu Gln Lys Ile Asp Leu Arg His Asn Glu Ile Tyr Glu Ile Lys Val					
1199	1208	1217	1226	1235	1244
GAC ACT TTC CAG CAG TTG CTT AGC CTC CGA TCG CTG AAT TTG GCT TGG AAC AAA					
---	---	---	---	---	---
Asp Thr Phe Gln Gln Leu Leu Ser Leu Arg Ser Leu Asn Leu Ala Trp Asn Lys					
1253	1262	1271	1280	1289	1298
ATT GCT ATT ATT CAC CCC AAT GCA TTT TCC ACT TTG CCA TCC CTA ATA AAG CTG					
---	---	---	---	---	---
Ile Ala Ile Ile His Pro Asn Ala Phe Ser Thr Leu Pro Ser Leu Ile Lys Leu					

FIGURE 1D

5/17

1307	1316	1325	1334	1343	1352
GAC CTA TCG TCC AAC CTC CTG TCG TCT TTT CCT ATA ACT GGG TTA CAT GGT TTA					
Asp Leu Ser Ser Asn Leu Leu Ser Ser Phe Pro Ile Thr Gly Leu His Gly Leu					
1361	1370	1379	1388	1397	1406
ACT CAC TTA AAA TTA ACA GGA AAT CAT GCC TTA CAG AGC TTG ATA TCA TCT GAA					
Thr His Leu Lys Leu Thr Gly Asn His Ala Leu Gln Ser Leu Ile Ser Ser Glu					
1415	1424	1433	1442	1451	1460
AAC TTT CCA GAA CTC AAG GTT ATA GAA ATG CCT TAT GCT TAC CAG TGC TGT GCA					
Asn Phe Pro Glu Leu Lys Val Ile Glu Met Pro Tyr Ala Tyr Gln Cys Cys Ala					
1469	1478	1487	1496	1505	1514
TTT GGA GTG TGT GAG AAT GCC TAT AAG ATT TCT AAT CAA TGG AAT AAA GGT GAC					
Phe Gly Val Cys Glu Asn Ala Tyr Lys Ile Ser Asn Gln Trp Asn Lys Gly Asp					
1523	1532	1541	1550	1559	1568
AAC AGC AGT ATG GAC GAC CTT CAT AAG AAA GAT GCT GGA ATG TTT CAG GCT CAA					
Asn Ser Ser Met Asp Asp Leu His Lys Lys Asp Ala Gly Met Phe Gln Ala Gln					
1577	1586	1595	1604	1613	1622
GAT GAA CGT GAC CTT GAA GAT TTC CTG CTT GAC TTT GAG GAA GAC CTG AAA GCC					
Asp Glu Arg Asp Leu Glu Asp Phe Leu Leu Asp Phe Glu Glu Asp Leu Lys Ala					

FIGURE 1E

6/17

1631	1640	1649	1658	1667	1676
CTT CAT TCA GTG CAG TGT TCA CCT TCC CCA GGC CCC TTC AAA CCC TGT GAA CAC					
---	---	---	---	---	---
Leu His Ser Val Gln Cys Ser Pro Ser Pro Gly Pro Phe Lys Pro Cys Glu His					
1685	1694	1703	1712	1721	1730
CTG CTT GAT GGC TGG CTG ATC AGA ATT GGA GTG TGG ACC ATA GCA GTT CTG GCA					
---	---	---	---	---	---
Leu Leu Asp Gly Trp Leu Ile Arg Ile Gly Val Trp Thr Ile Ala Val Leu Ala					
1739	1748	1757	1766	1775	1784
CTT ACT TGT AAT GCT TTG GTG ACT TCA ACA GTT TTC AGA TCC CCT CTG TAC ATT					
---	---	---	---	---	---
Leu Thr Cys Asn Ala Leu Val Thr Ser Thr Val Phe Arg Ser Pro Leu Tyr Ile					
1793	1802	1811	1820	1829	1838
TCC CCC ATT AAA CTG TTA ATT GGG GTC ATC GCA GCA GTG AAC ATG CTC ACG GGA					
---	---	---	---	---	---
Ser Pro Ile Lys Leu Leu Ile Gly Val Ile Ala Ala Val Asn Met Leu Thr Gly					
1847	1856	1865	1874	1883	1892
GTC TCC AGT GCC GTG CTG GCT GGT GTG GAT GCG TTC ACT TTT GGC AGC TTT GCA					
---	---	---	---	---	---
Val Ser Ser Ala Val Leu Ala Gly Val Asp Ala Phe Thr Phe Gly Ser Phe Ala					
1901	1910	1919	1928	1937	1946
CGA CAT GGT GCC TGG TGG GAG AAT GGG GTT GGT TGC CAT GTC ATT GGT TTT TTG					
---	---	---	---	---	---
Arg His Gly Ala Trp Trp Glu Asn Gly Val Gly Cys His Val Ile Gly Phe Leu					

FIGURE 1F

7/17

1955	1964	1973	1982	1991	2000
TCC ATT TTT	GCT TCA GAA TCA	TCT GTT TTC	CTG CTT ACT	CTG GCA GCC	CTG GAG
---	---	---	---	---	---
Ser Ile Phe	Ala Ser Glu Ser	Val Ser Val Phe	Leu Leu Thr	Leu Ala Ala	Leu Glu
2009	2018	2027	2036	2045	2054
CGT GGG TTC	TCT GTG AAA TAT	TCT GCA AAA	TTT GAA ACG	AAA GCT CCA	TTT TCT
---	---	---	---	---	---
Arg Gly Phe	Ser Val Lys Tyr	Ser Ala Lys	Phe Glu Thr	Lys Ala Pro	Phe Ser
2063	2072	2081	2090	2099	2108
AGC CTG AAA	GTA ATC ATT TTG	CTC TGT GCC	CTG CTG GCC	TTG ACC ATG	GCC GCA
---	---	---	---	---	---
Ser Leu Lys	Val Ile Ile Leu	Cys Ala Leu	Leu Ala Leu	Thr Met Ala	Ala
2117	2126	2135	2144	2153	2162
GTT CCC CTG	CTG GGT GGC AGC	AAG TAT GGC	GCC TCC CCT	CTC TGC CTG	CCT TTG
---	---	---	---	---	---
Val Pro Leu	Leu Gly Gly Ser	Lys Tyr Gly	Ala Ser Pro	Leu Cys Leu	Pro Leu
2171	2180	2189	2198	2207	2216
CCT TTT GGG	GAG CCC AGC ACC	ATG GGC TAC	ATG GTC GCT	CTC ATC TTG	CTC AAT
---	---	---	---	---	---
Pro Phe Gly	Glu Pro Ser Thr	Met Gly Tyr	Met Val Ala	Ile Leu Leu	Asn
2225	2234	2243	2252	2261	2270
TCC CTT TGC	TTC CTC ATG ATG	ACC ATT GCC	TAC ACC AAG	CTC TAC TGC	AAT TTG
---	---	---	---	---	---
Ser Leu Cys	Phe Leu Met Met	Thr Ile Ala Tyr	Thr Thr Lys	Leu Tyr Cys	Asn Leu

FIGURE 1G



8/17

2279	2288	2297	2306	2315	2324
GAC AAG GGA GAC CTG GAG AAT ATT TGG GAC TGC TCT ATG GTA AAA CAC ATT GCC					
---	---	---	---	---	---
Asp Lys Gly Asp Leu Glu Asn Ile Trp Asp Cys Ser Met Val Lys His Ile Ala					
2333	2342	2351	2360	2369	2378
CTG TTG CTC TTC ACC AAC TGC ATC CTA AAC TGC CCT GTG GCT TTC TTG TCC TTC					
---	---	---	---	---	---
Leu Leu Leu Phe Thr Asn Cys Ile Leu Asn Cys Pro Val Ala Phe Leu Ser Phe					
2387	2396	2405	2414	2423	2432
TCC TCT TTA ATA AAC CTT ACA TTT ATC AGT CCT GAA GTA ATT AAG TTT ATC CTT					
---	---	---	---	---	---
Ser Ser Leu Ile Asn Leu Thr Phe Ile Ser Pro Glu Val Ile Lys Phe Ile Leu					
2441	2450	2459	2468	2477	2486
CTG GTG GTA GTC CCA CTT CCT GCA TGT CTC AAT CCC CTT CTC TAC ATC TTG TTC					
---	---	---	---	---	---
Leu Val Val Val Pro Leu Pro Ala Cys Leu Asn Pro Leu Leu Tyr Ile Leu Phe					
2495	2504	2513	2522	2531	2540
AAT CCT CAC TTT AAG GAG GAT CTG GTG AGC CTG AGA AAG CAA ACC TAC TTC TGG					
---	---	---	---	---	---
Asn Pro His Phe Lys Glu Asp Leu Val Ser Leu Arg Lys Gln Thr Tyr Val Trp					
2549	2558	2567	2576	2585	2594
ACA AGA TCA AAA CAC CCA AGC TTG ATG TCA ATT AAC TCT GAT GTC GAA AAA					
---	---	---	---	---	---
Thr Arg Ser Lys His Pro Ser Leu Met Ser Ile Asn Ser Asp Val Glu Lys					

FIGURE 1H

9/17

2603	2612	2621	2630	2639	2648
CAG TCC TGT GAC TCA ACT CAA GCC TTG GTA ACC TTT ACC AGC TCC AGC ATC ACT					
---	---	---	---	---	---
Gln Ser Cys Asp Ser Thr Gln Ala Leu Val Thr Phe Thr Ser Ser Ile Thr					
2657	2666	2675	2684	2693	2702
TAT GAC CTG CCT CCC AGT TCC GTG CCA TCA CCA GCT TAT CCA GTG ACT GAG AGC					
---	---	---	---	---	---
Tyr Asp Leu Pro Pro Ser Ser Val Pro Ser Pro Ala Tyr Pro Val Thr Glu Ser					
2711	2720	2729	2738	2747	2756
TGC CAT CTT TCC TCT GTG GCA TTT GTC CCA TGT CTC TAA TTA ATA TGT GAA GGA					
---	---	---	---	---	---
Cys His Leu Ser Ser Val Ala Phe Val Pro Cys Leu					
2765	2774	2783	2792	2801	2810
AAA TGT TTT CAA AGG TTG AGA ACC TGA AAA TGT GAG ATT GAG TAT ATC AGA GCA					
---	---	---	---	---	---

FIGURE 11

10/17

2819	2828	2837	2846	2855	2864
GTA ATT AAT AAG AAG AGC TGA GGT GAA ACT CGG TTT AAA AAC CAA AAA AGA ATC					
---	---	---	---	---	---
2873	2882	2891	2900	2909	2918
TCT CAG TTA GTA AGA AAA GGC TGA AAA CCT CTT GAT ACT TGA GAG TGA ATA TAA					
---	---	---	---	---	---
2927	2936	2945	2954	2963	2972
GTC TAA ATG CTG CTT TGT ATA ATT TGT TCA GCT AAG GGA TAG ATC GAT CAC ACT					
---	---	---	---	---	---

A 3'

FIGURE 1J

11/17

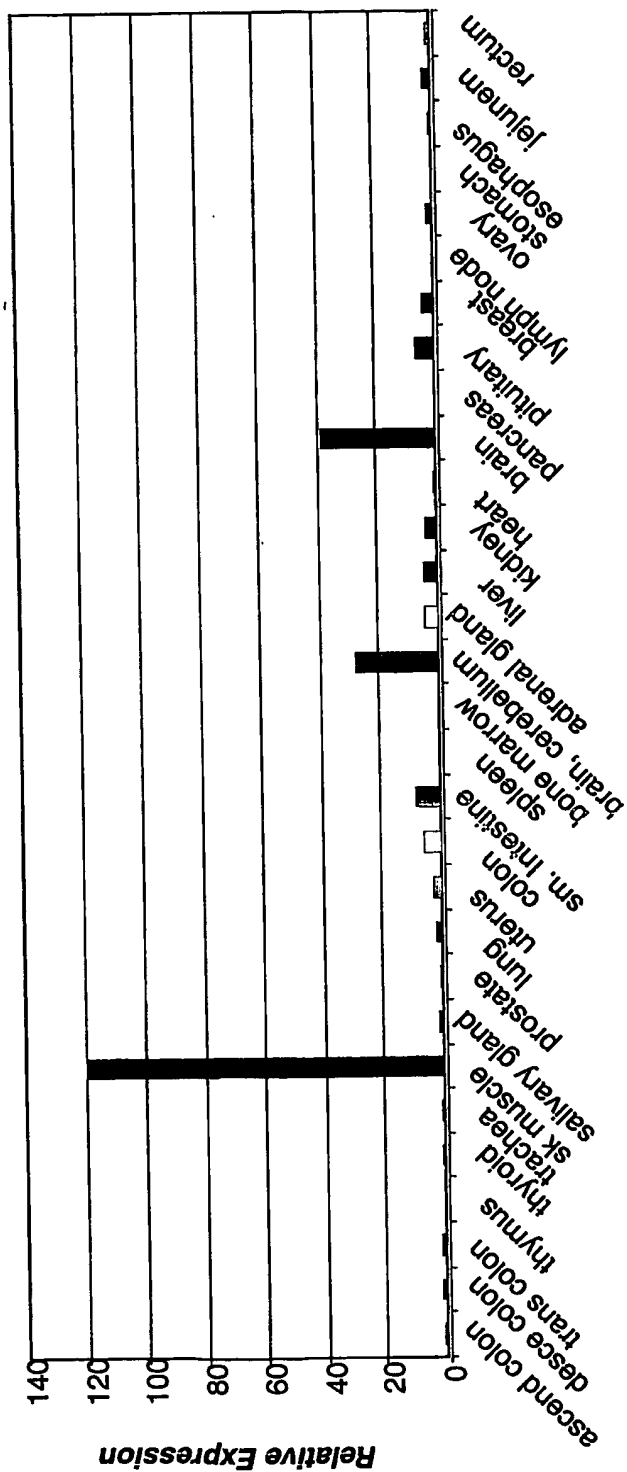


FIGURE 2

12/17

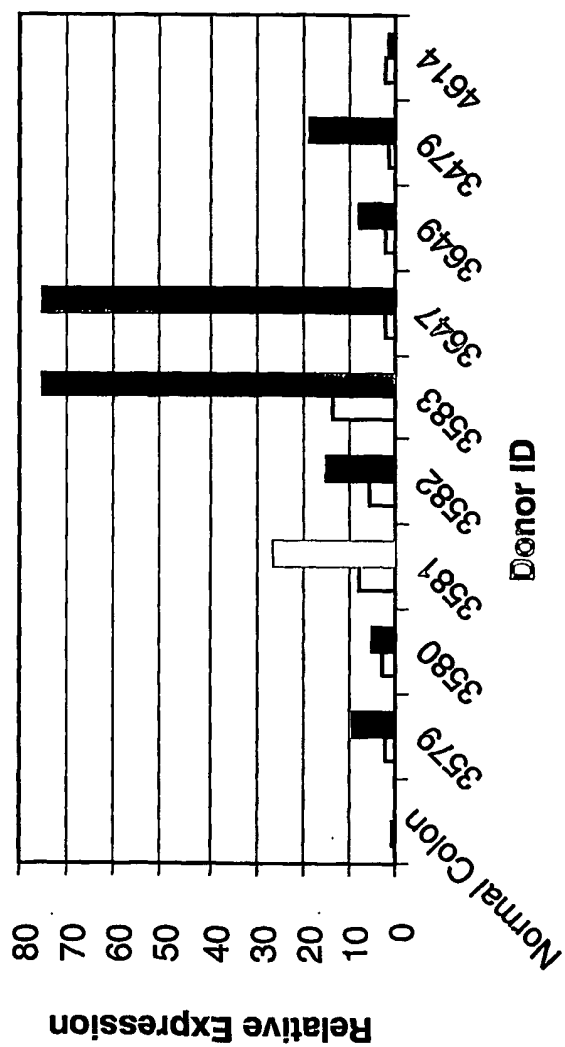


FIGURE 3

13/17

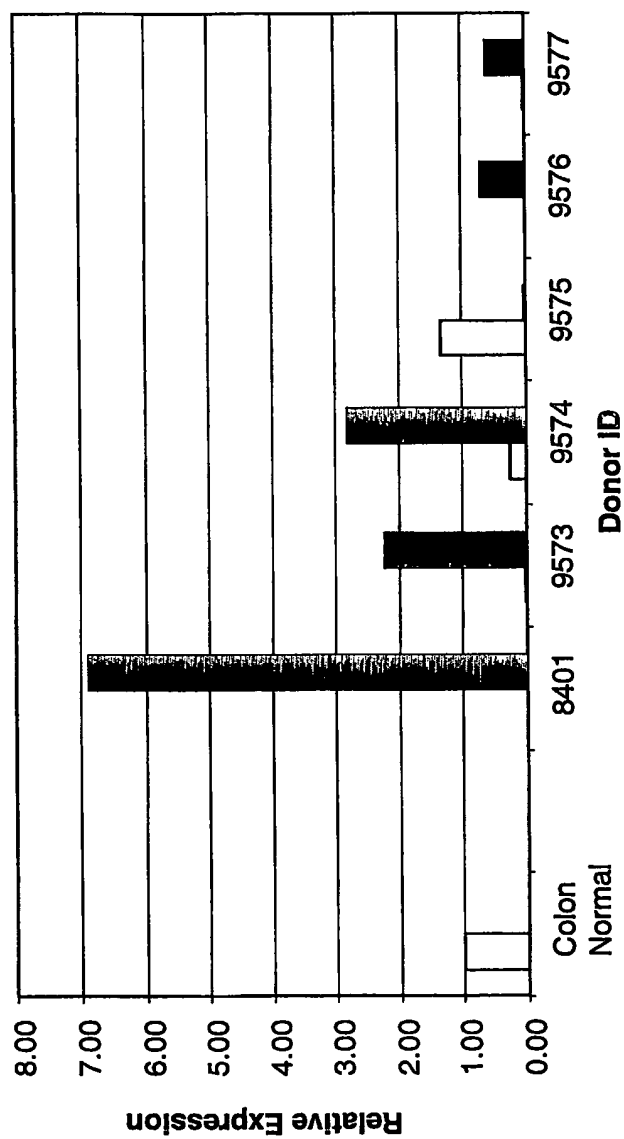


FIGURE 4

14/17

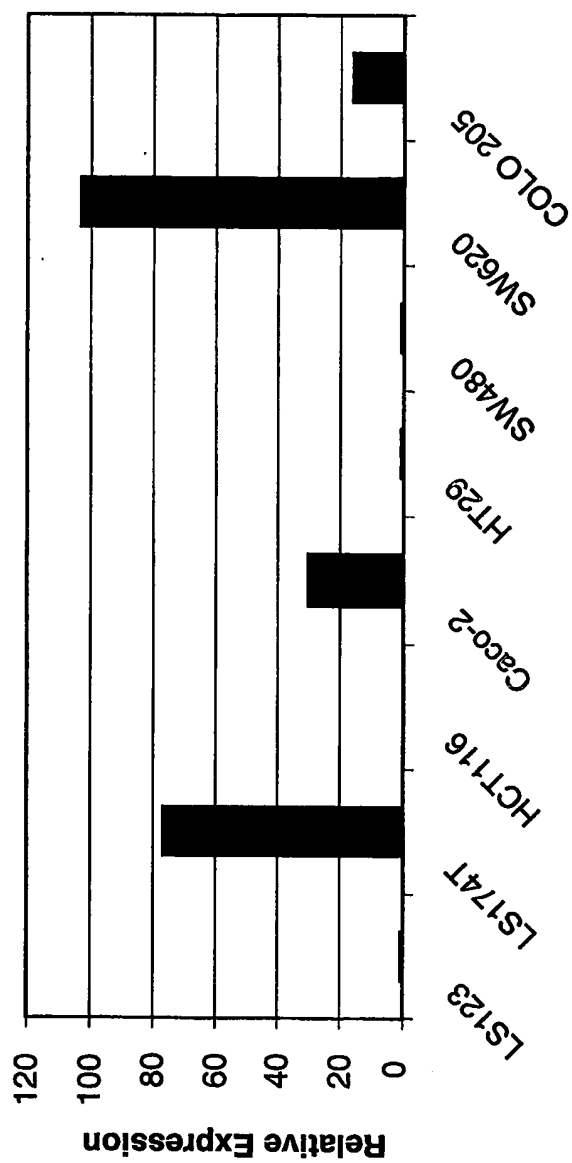


FIGURE 5

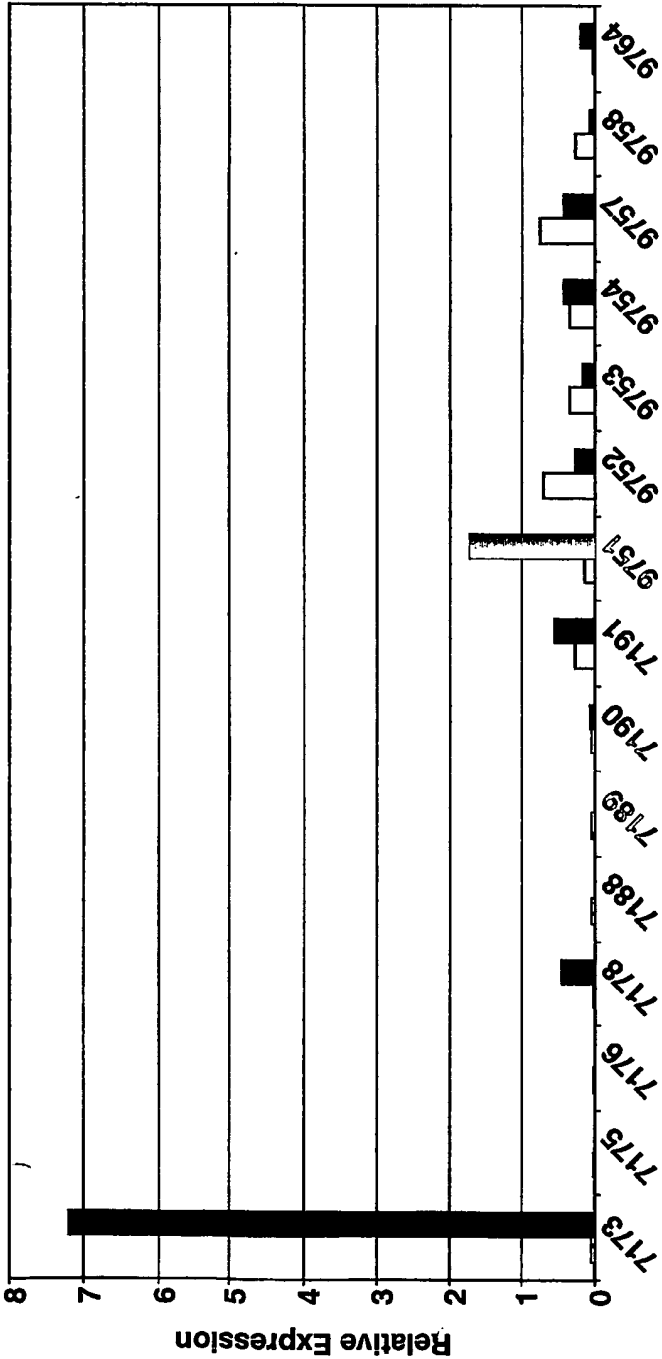


FIGURE 6



16/17

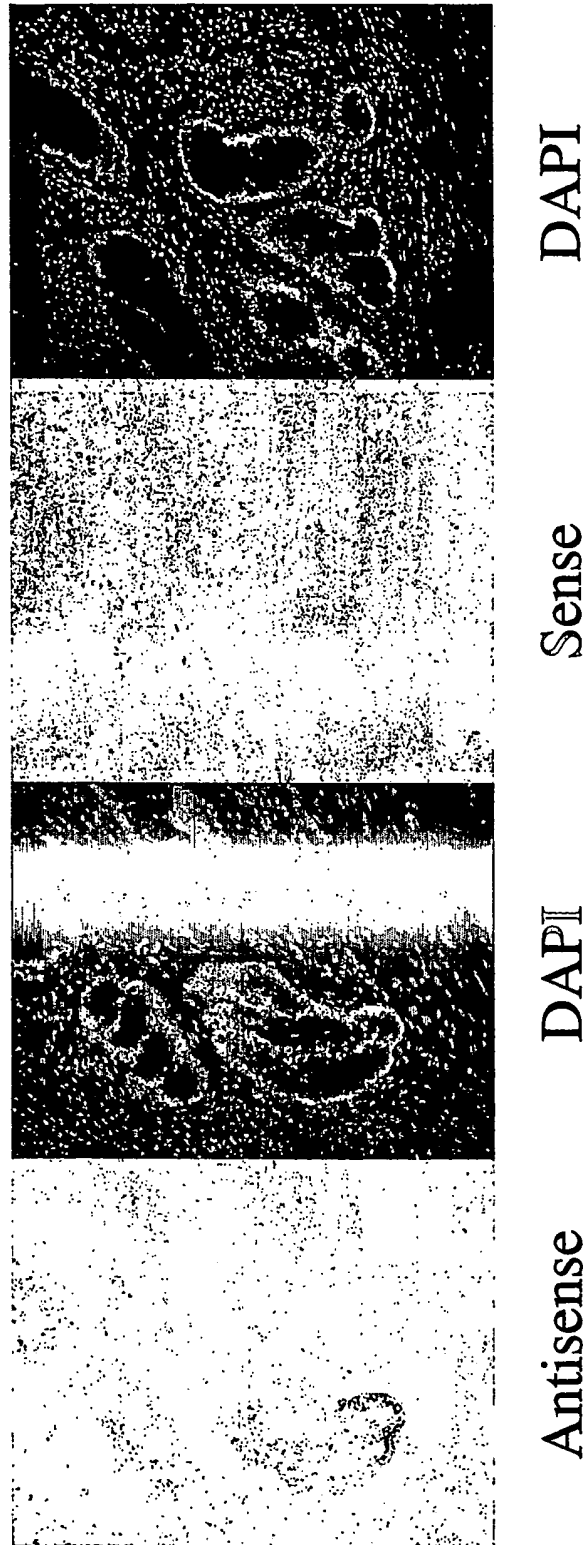


FIGURE 7

17/17

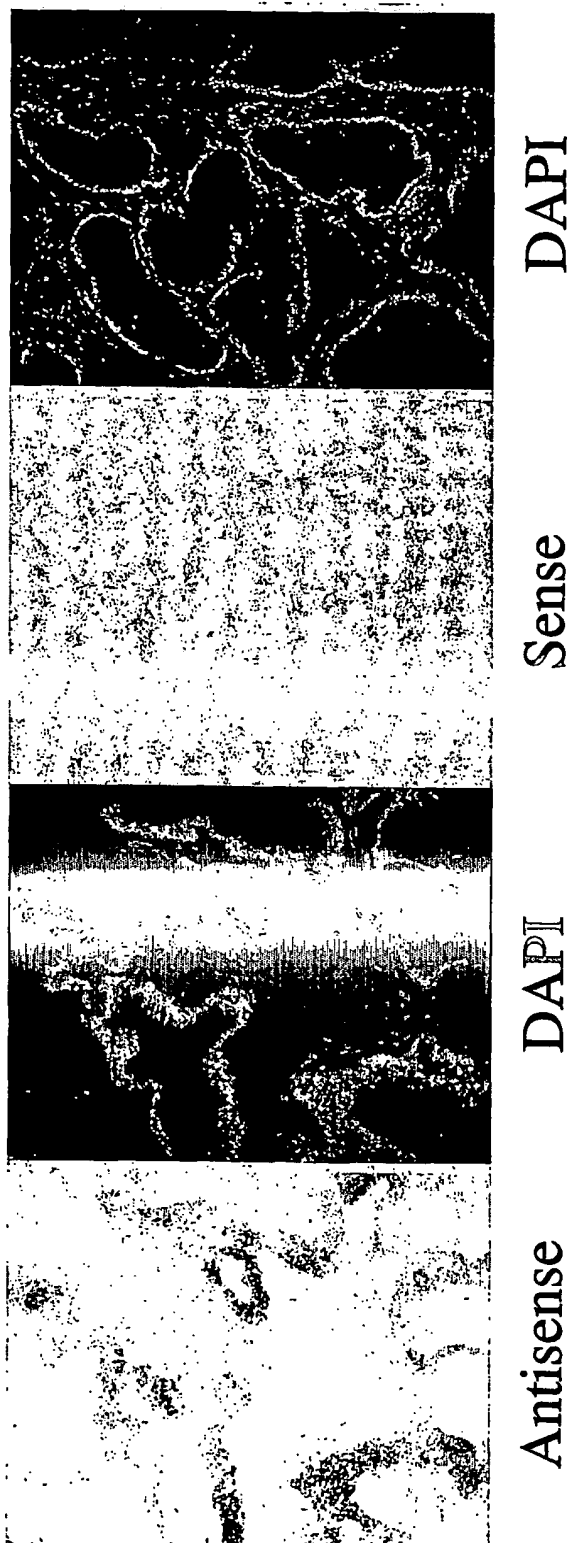


FIGURE 8



Ser	Ile	Pro	Glu	Lys	Ala	Phe	Val	Gly	Asn	Pro	Ser	Leu	Ile	Thr
				275					280					285
Ile	His	Phe	Tyr	Asp	Asn	Pro	Ile	Gln	Phe	Val	Gly	Arg	Ser	Ala
				290					295					300
Phe	Gln	His	Leu	Pro	Glu	Leu	Arg	Thr	Leu	Thr	Leu	Asn	Gly	Ala
				305					310					315
Ser	Gln	Ile	Thr	Glu	Phe	Pro	Asp	Leu	Thr	Gly	Thr	Ala	Asn	Leu
				320					325					330
Glu	Ser	Leu	Thr	Leu	Thr	Gly	Ala	Gln	Ile	Ser	Ser	Leu	Pro	Gln
				335					340					345
Thr	Val	Cys	Asn	Gln	Leu	Pro	Asn	Leu	Gln	Val	Leu	Asp	Leu	Ser
				350					355					360
Tyr	Asn	Leu	Leu	Glu	Asp	Leu	Pro	Ser	Phe	Ser	Val	Cys	Gln	Lys
				365					370					375
Leu	Gln	Lys	Ile	Asp	Leu	Arg	His	Asn	Glu	Ile	Tyr	Glu	Ile	Lys
				380					385					390
Val	Asp	Thr	Phe	Gln	Gln	Leu	Leu	Ser	Leu	Arg	Ser	Leu	Asn	Leu
				395					400					405
Ala	Trp	Asn	Lys	Ile	Ala	Ile	Ile	His	Pro	Asn	Ala	Phe	Ser	Thr
				410					415					420
Leu	Pro	Ser	Leu	Ile	Lys	Leu	Asp	Leu	Ser	Ser	Asn	Leu	Leu	Ser
				425					430					435
Ser	Phe	Pro	Ile	Thr	Gly	Leu	His	Gly	Leu	Thr	His	Leu	Lys	Leu
				440					445					450
Thr	Gly	Asn	His	Ala	Leu	Gln	Ser	Leu	Ile	Ser	Ser	Glu	Asn	Phe
				455					460					465
Pro	Glu	Leu	Lys	Val	Ile	Glu	Met	Pro	Tyr	Ala	Tyr	Gln	Cys	Cys
				470					475					480
Ala	Phe	Gly	Val	Cys	Glu	Asn	Ala	Tyr	Lys	Ile	Ser	Asn	Gln	Trp
				485					490					495
Asn	Lys	Gly	Asp	Asn	Ser	Ser	Met	Asp	Asp	Leu	His	Lys	Lys	Asp
				500					505					510
Ala	Gly	Met	Phe	Gln	Ala	Gln	Asp	Glu	Arg	Asp	Leu	Glu	Asp	Phe
				515					520					525
Leu	Leu	Asp	Phe	Glu	Glu	Asp	Leu	Lys	Ala	Leu	His	Ser	Val	Gln
				530					535					540
Cys	Ser	Pro	Ser	Pro	Gly	Pro	Phe	Lys	Pro	Cys	Glu	His	Leu	Leu
				545					550					555
Asp	Gly	Trp	Leu	Ile	Arg	Ile	Gly	Val	Trp	Thr	Ile	Ala	Val	Leu
				560					565					570
Ala	Leu	Thr	Cys	Asn	Ala	Leu	Val	Thr	Ser	Thr	Val	Phe	Arg	Ser
				575					580					585
Pro	Leu	Tyr	Ile	Ser	Pro	Ile	Lys	Leu	Leu	Ile	Gly	Val	Ile	Ala
				590					595					600
Ala	Val	Asn	Met	Leu	Thr	Gly	Val	Ser	Ser	Ala	Val	Leu	Ala	Gly
				605					610					615
Val	Asp	Ala	Phe	Thr	Phe	Gly	Ser	Phe	Ala	Arg	His	Gly	Ala	Trp
				620					625					630
Trp	Glu	Asn	Gly	Val	Gly	Cys	His	Val	Ile	Gly	Phe	Leu	Ser	Ile
				635					640					645
Phe	Ala	Ser	Glu	Ser	Ser	Val	Phe	Leu	Leu	Thr	Leu	Ala	Ala	Leu
				650					655					660
Glu	Arg	Gly	Phe	Ser	Val	Lys	Tyr	Ser	Ala	Lys	Phe	Glu	Thr	Lys
				665					670					675
Ala	Pro	Phe	Ser	Ser	Leu	Lys	Val	Ile	Ile	Leu	Leu	Cys	Ala	Leu
				680					685					690
Leu	Ala	Leu	Thr	Met	Ala	Ala	Val	Pro	Leu	Leu	Gly	Gly	Ser	Lys
				695					700					705
Tyr	Gly	Ala	Ser	Pro	Leu	Cys	Leu	Pro	Leu	Pro	Phe	Gly	Glu	Pro
				710					715					720
Ser	Thr	Met	Gly	Tyr	Met	Val	Ala	Leu	Ile	Leu	Leu	Asn	Ser	Leu
				725					730					735
Cys	Phe	Leu	Met	Met	Thr	Ile	Ala	Tyr	Thr	Lys	Leu	Tyr	Cys	Asn

	740		745		750
Leu Asp Lys Gly	Asp Leu Glu Asn Ile	Trp Asp Cys Ser Met	Val		
	755		760		765
Lys His Ile Ala	Leu Leu Leu Phe Thr	Asn Cys Ile Leu Asn	Cys		
	770		775		780
Pro Val Ala Phe	Leu Ser Phe Ser Ser	Leu Ile Asn Leu Thr	Phe		
	785		790		795
Ile Ser Pro Glu	Val Ile Lys Phe Ile	Leu Leu Val Val Val	Pro		
	800		805		810
Leu Pro Ala Cys	Leu Asn Pro Leu Leu	Tyr Ile Leu Phe Asn	Pro		
	815		820		825
His Phe Lys Glu	Asp Leu Val Ser Leu	Arg Lys Gln Thr Tyr	Val		
	830		835		840
Trp Thr Arg Ser	Lys His Pro Ser Leu	Met Ser Ile Asn Ser	Asp		
	845		850		855
Asp Val Glu Lys	Gln Ser Cys Asp Ser	Thr Gln Ala Leu Val	Thr		
	860		865		870
Phe Thr Ser Ser	Ser Ile Thr Tyr Asp	Leu Pro Pro Ser Ser	Val		
	875		880		885
Pro Ser Pro Ala	Tyr Pro Val Thr Glu	Ser Cys His Leu Ser	Ser		
	890		895		900
Val Ala Phe Val	Pro Cys Leu				
	905				

&lt;210&gt; 2

&lt;211&gt; 2973

&lt;212&gt; DNA

&lt;213&gt; Homo sapiens

&lt;220&gt;

&lt;221&gt; misc\_feature

&lt;223&gt; Incyte ID No: 1736415CB1

&lt;400&gt; 2

```

ccctacttc gggcaccatg gacacctccc ggctcgggtg gtcctgtgcc ttgcctgtgc 60
tgctgcagct ggcgaccggg ggcagctctc ccaggctctg tgtgttgctg aggggctgcc 120
ccacacactg tcattgcgag cccgacggca ggatgttgct caggggtggac tgctccgacc 180
tggggctctc ggagctgcct tccaacctca gcgtcttcac ctcctaccta gacctcagta 240
tgaacaacat cagtcagctg ctccccgaatc ccctgcccag tctccgcttc ctggaggagt 300
tacgtcttgc gggaaacgct ctgacataca ttcccaaggg agcattcact ggcctttaca 360
gtcttaagtg tcttatgctg cagaataatc agctaagaca cgtaccacac gaagctctgc 420
agaatttgcg aagccttcaa tccctgcgtc tggatgctaa ccacatcagc tatgtgcccc 480
caagctgttt cagtggcctg cattccctga ggcacctgtg gctggatgac aatgcgttaa 540
cagaaatccc cgtccaggct tttagaagtt tatcggcatt gcaagccatg accttgcccc 600
tgaacaaaat acaccacata ccagactatg cctttggaaa cctctccagc ttggtagttc 660
tacatctcca taacaataga atccactccc tgggaaagaa atgctttgat gggctccaca 720
gcctagagac tttagattta aattacaata accttgatga attccccact gcaattagga 780
cactctccaa ccttaaagaa ctaggatttc atagcaacaa tatcaggtcg atacctgaga 840
aagcatttgt aggcaaccct tctcttatta caatacattt ctatgacaat cccatccaat 900
ttgttgggag atctgctttt caacatttac ctgaactaag aacactgact ctgaatgggtg 960
cctcacaaat aactgaattt cctgatttaa ctggaactgc aaacctggag agtctgactt 1020
taactggagc acagatctca tctcttcttc aaaccgtctg caatcagtta cctaactctc 1080
aagtgcctga tctgtcttac aacctattag aagatttacc cagtttttca gtctgcaaaa 1140
agcttcagaa aattgacctc agacataatg aaatctacga aattaaagt gacactttcc 1200
agcagttgct tagcctccga tcgctgaatt tggtttggaa caaaattgct attattcacc 1260
ccaatgcatt ttccactttg ccatccctaa taaagctgga cctatcgtcc aacctcctgt 1320
cgtcttttcc tataactggg ttacatgggt taactcactt aaaattaaca ggaaatcatg 1380
ccttacagag cttgatatca tctgaaaact ttccagaact caaggttata gaaatgcctt 1440
atgcttacca gtgctgtgca tttggagtgt gtgagaatgc ctataagatt tctaataaat 1500
ggaataaagg tgacaacagc agtatggagc accttcataa gaaagatgct ggaatgtttc 1560
aggctcaaga tgaacgtgac cttgaagatt tcctgcttga ctttgaggaa gacctgaaag 1620
cccttcattc agtgcagtgt tcaccttccc caggccccctt caaacctgtg gaacacctgc 1680
ttgatggctg gctgatcaga attggagtgt ggaccatagc agttctggca cttacttgta 1740

```

```

atgcttttgggt gacttcaaca gttttcagat cccctctgta catttcccc attaaactgt 1800
taattgggggt catcgagca gtgaacatgc tcacgggagt ctccagtgcc gtgctggctg 1860
gtgtggatgca gttcactttt ggcagctttg cacgacatgg tgcctgggtg gagaatgggg 1920
ttgggttgcca tgtcattgggt tttttgtcca tttttgcttc agaatacatct gttttcctgc 1980
ttactctggc agccctggag cgtgggttct ctgtgaaata ttctgcaaaa tttgaaacga 2040
aagctccatt ttctagcctg aaagtaatca ttttgcctg tgcctgctg gccttgacca 2100
tggcgcagct tccctgctg ggtggcagca agtatggcgc ctccctctc tgcctgcctt 2160
tgccttttgg ggagcccagc accatgggct acatggctgc tctcatcttg ctcaattccc 2220
tttgcttcct catgatgacc attgcctaca ccaagctcta ctgcaatttg gacaagggag 2280
acctggagaa tatttgggac tgccttatgg taaaacacat tgcctgttg ctcttcacca 2340
actgcatcct aaactgccct gtggctttct tgccttctc ctcttaata aaccttacct 2400
ttatcagtc tgaagtaatt aagtttatcc ttctgggtgt agtccactt cctgcatgtc 2460
tcaatccct tctctacatc ttgttcaatc ctcactttaa ggaggatctg gtgagcctga 2520
gaaagcaaac ctacgtctgg acaagatcaa aacaccaag ctgatgtca ataacctctg 2580
atgatgtcga aaaacagtc tgtgactcaa ctcaagcctt ggtaaccttt accagctcca 2640
gcatacctta tgacctgcct cccagttcgc tgcctacc agcttatcca gtgactgaga 2700
gctgccatct ttctctgtg gcatttgtcc catgtctcta ataatatgt gaaggaaaat 2760
gttttcaaag gttgagaacc tgaatatgtg agattgagta tatcagagca gtaattaata 2820
agaagagctg aggtgaaact cggtttaaaa accaaaaag aatctctcag ttagtaagaa 2880
aaggctgaaa acctcttgat acttgagagt gaatataagt ctaaatgctg ctttgtataa 2940
tttgttcagc taagggatag atcgatcaca cta 2973

```

&lt;210&gt; 3

&lt;211&gt; 3098

&lt;212&gt; DNA

&lt;213&gt; Mus musculus

&lt;220&gt;

&lt;221&gt; misc\_feature

&lt;223&gt; Incyte ID No: 050237\_Mm.1

&lt;400&gt; 3

```

cttaaaactc taacctttta aacaaggaat ttgtggattt aaatacacc attgagtcac 60
ccaacgagtc ttctcactat gggagactat gctgggtgagt cagaagagac tacaatgcga 120
agagcatttc aggtctgttt tctcatctgg gttcacctaa gtgacttgat ggctttcctt 180
cttagctgaa caaatgatac aagcgggtgt taatgatgct ctccgtctct aaggcaccac 240
tggtttccaca ccaatttaga gatgtttgct tgggttttaa ggtaggtttt tagctcagct 300
tttcttagct actgctctga tagaaatcac ttttcagggt tccaactctt aaaaacgttc 360
ctctctcata gtcactagag acatgggaca aatgcaactg aagagagatg acagctttca 420
gtcatggggt aagctgggtg tgcctcgga gtggaaggca agtcataaggc tatgctggcg 480
tgggtaaagg atactaaggc ttgggttgac tcacaggacc gtttctcaac atcgctccag 540
ttaatggaca gcagactcgc gtgttttgat ctcatccaga aacgggtatg ctttccagg 600
ctgcccatac cctccttaaa atggggattg aagacaatgt agagaagtgg gttgagacag 660
gaaggaagtg ggacgatcac gagaagtata aatttaatac cgtcaggact gataaagggt 720
aggttttagc aagaggagaa ggataagaaa gccacggggc agtaaaggat gcagttggcg 780
aagagcaaca gagcaatgtg cttcaccatc gaacaatccc aaagattctc cagctctcct 840
ttctccaaac tgcagtagag ctttgtgtag gcaatggtca ttatgaggaa acagagagag 900
ttgagcaaca cgagagccac catgtagccc gtggtgctgg gctcccaaa gggcaagggc 960
aggcagaggg gagaggcatt gtacttactg cctcctagca aggggtattg ggcaatggtc 1020
agggccaaac ggacacatag caaaacgatc gctctcaggc taaaaagggg agctttcact 1080
tcaaacttcg aagagcactt gacagaaaaa cctcgttcca gcgtgcccag agtgagcagg 1140
aagatcgacg attcggaagc aaaaatggac aggaagccaa cgatttggca gccgattccg 1200
tcttcccacc acgcaccgtg ctgagcaaaa cggccaaaag tgaatgcac cacggcagcc 1260
agcacagcac tggagacccc catgagaatg tccactaccg cgattacccc aattagcagc 1320
tttatggaag agatgtacag gggagttctg aacacggtca aagccaccaa ggcattgcag 1380
gaaagcgtca gtactgccgt ggtccacacc cgcattcgga tcagccagct accaaatagg 1440
tgctcacagg gcttgaaggg acctggggaa gcgcagcact gcaccgagtg aagggcattc 1500
aggtcttctc caaagtcaag taggaaatct tcaaggtccc gctcatcttg aacttgaaat 1560
aaccagcgt ctttcttatg aaggtcgtcc acactgttgc cgtcgtcttt attccattgg 1620
ttagaaatth tatagacatt ctacaccccc ccaaatgcac aacactggta agcagatggc 1680
tttctataa tcttgagctc tgggaagtth tcagatggta tcaggctctg taaggctcgg 1740
ttccctgtta attttaagtg agttaaacca tgtaaccagg tcacagggaa ggacgacagg 1800
agattggatg atagggtcaa ctttattaga gacggcaacg tagaaaaacgc attgggggtg 1860

```

atgatagcaa	ttttattcca	tgctaagttc	agagatcgga	ggtaaaca	ctgctgaaa	1920
gtgctgccct	taatttcata	gatctcgta	tgctcaggt	caattttctg	aagtttttgg	1980
cagcctgaca	aactgggtaa	gtcttcgagt	aggttgtaag	acaaatctag	cacttggaga	2040
ttaggtaact	gatcacagac	ggcctgggga	agagatgaga	tctttgctcc	agttaaagtc	2100
agactctcca	gggtggcagt	tcctgtcaag	tgaggaaatt	cagtaatgtg	cgaggcacca	2160
ttcaaagtca	gtgttcttag	ttcaggcaaa	tgctgaaaag	cagatactcc	aacaaattgg	2220
atggggttgt	catagaagtg	tattgtgata	agagaagggg	tgctacgaa	cgctcgctcc	2280
gggtattgacc	tgatgttggt	gctgtggaat	cctagttcct	taaggttgga	gagtgtcttg	2340
attgcagtgg	ggaattcatc	aaggttatta	taatttaa	ctaaagtctc	caggctgtgg	2400
agtcacatcaa	agcatttctt	tcccagggag	tggtattctat	tattatggag	atgcagaacc	2460
acgaggctgg	agaggtttcc	aaaggcgtag	tctgctatgt	gggtattttt	gttcagggcc	2520
aagggtcatgg	cttgcagggc	tgataaaactt	ctgaaagcct	ggacagggac	gtctgtgaga	2580
gcattgtcat	ctagccacag	gtgcctcagg	gagtgcaggc	cgctgaaaca	gctgggtggc	2640
acgtagctga	tgtgggtggc	atctaggcgc	agggattgaa	ggcttctcaa	attctgtagc	2700
gcttccctcg	gaacctttct	cagctgggtg	ttctgcagca	taagcacttt	gaggctgtga	2760
aggcccgtag	acgctccctt	gggaatgtgt	gtcaaagcat	ttccagcaag	acgtaactct	2820
tctaggaagc	agaggcgatg	taggagactg	gcgggtagct	gactgatgtt	gttcatactg	2880
aggtccaggt	aggaggtgaa	gacgctgagg	ttggagggca	gctccgagag	ccccaggtcc	2940
gagcagctcca	ccctgagcag	catcctgcca	tccagctcac	agtgacagtg	tgatgggcag	3000
ccccgcggtg	tcgcatctgg	tcccgggtgag	ctgccggcgg	ccaccaactg	cagcagcgcc	3060
agcaaggaca	ggagcatgtg	gacgcaggag	gtgtccat			3098